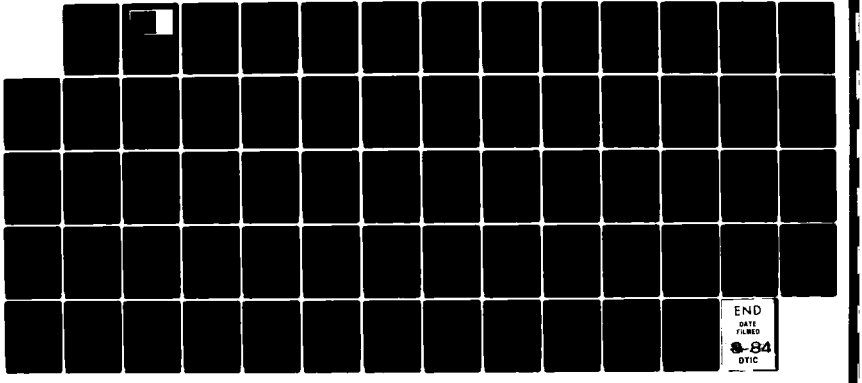
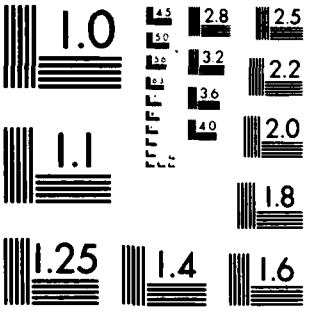
AD-A142 768 BAYESIAN MODELS FOR RESPONSE SURFACES II ESTIMATING THE 1/
RESPONSE SURFACE(U) WISCONSIN UNIV-MADISON MATHEMATICS
RESEARCH CENTER D M STEINBERG APR 84 MRC-TSR-2683
UNCLASSIFIED DAAG29-80-C-0041 F/G 12/1 NL





MRC Technical Summary Report #2683

BAYESIAN MODELS FOR RESPONSE SURFACES II: ESTIMATING THE RESPONSE SURFACE

David M. Steinberg

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53705

April 1984

(Received April 16, 1984)

Approved for public release
Distribution unlimited

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

National Science Foundation
Washington, D. C. 20550

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: 2683	2. GOVT ACCESSION NO.: AD-A142 768	3. RECIPIENT'S CATALOG NUMBER:
4. TITLE (and Subtitle): BAYESIAN MODELS FOR RESPONSE SURFACES II: ESTIMATING THE RESPONSE SURFACE		5. TYPE OF REPORT & PERIOD COVERED: Summary Report - no specific reporting period
		6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): David M. Steinberg		8. CONTRACT OR GRANT NUMBER(s): DAAG29-80-C-0041 MCS-8210950
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Mathematics Research Center, University of Wisconsin 610 Walnut Street Madison, Wisconsin 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: Work Unit Number 4 - Statistics and Probability
11. CONTROLLING OFFICE NAME AND ADDRESS: See Item 18 below.		12. REPORT DATE: April 1984
		13. NUMBER OF PAGES: 61
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report): UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

National Science Foundation
Washington, D. C. 20550

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Response Surface Models
Bayesian Estimation
Smoothing Splines
Linear Regression
Polynomial Regression
Hierarchical Linear Model
Bayesian Linear Model

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

We analyze a Bayesian model for response surfaces. This model augments a simple graduating function with a bias term that represents the difference between the true, but unknown, response function and the graduating function. The model is a straightforward extension of Blight and Ott's (1975) Bayesian model for polynomial regression. The bias term is defined in terms of prior

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. ABSTRACT - cont'd.

distributions and we show how estimates of the response surface and measures of precision depend on the form of the prior distribution. Estimates and measures of precision are given for strictly proper prior distributions and also when improper priors are assigned to the coefficients of the graduating function, in which case the model leads to generalized spline estimates.

UNIVERSITY OF WISCONSIN-MADISON
MATHEMATICS RESEARCH CENTER

BAYESIAN MODELS FOR RESPONSE SURFACES II: ESTIMATING THE RESPONSE SURFACE*

David M. Steinberg

Technical Summary Report #2683
April 1984

ABSTRACT

We analyze a Bayesian model for response surfaces. This model augments a simple graduating function with a bias term that represents the difference between the true, but unknown, response function and the graduating function. The model is a straightforward extension of Blight and Ott's (1975) Bayesian model for polynomial regression. The bias term is defined in terms of prior distributions and we show how estimates of the response surface and measures of precision depend on the form of the prior distribution. Estimates and measures of precision are given for strictly proper prior distributions and also when improper priors are assigned to the coefficients of the graduating function, in which case the model leads to generalized spline estimates.

AMS (MOS) Subject Classifications: 62F15, 62J05

Key Words: Response Surface Models; Bayesian Estimation; Smoothing Splines; Linear Regression; Polynomial Regression; Hierarchical Linear Model; Bayesian Linear Model

Work Unit Number 4 (Statistics and Probability)

---

*This work forms part of the author's Ph. D. dissertation written under the direction of Professor G. E. P. Box. The author is grateful to Professor Box for his many valuable comments. The author would also like to thank Professor R. Myers for permission to use the data cited in Section 8.

This research was sponsored by the United States Army under Contract No. DAAG29-80-C-0041. This material is based upon work supported by the National Science Foundation under Grant No. MCS-8210950.

SIGNIFICANCE AND EXPLANATION

Scientists often wish to describe the relationship between a response variable and a collection of explanatory variables. When the particular nature of the relationship is unknown, as is often the case, a common strategy is to develop an empirical model by using a simple graduating function such as a low-degree polynomial to approximate the true relationship. The techniques of response surface methodology were developed to accomplish this goal.

We consider a generalization of standard response surface methodolgy that attempts to take into account the approximate nature of the graduating functions that are used. We propose a model in which the graduating function is augmented by a bias term that represents the difference between the true, but unknown, response function and the graduating function. The bias term is characterized in terms of the scientist's prior beliefs about its likely magnitude and its likely similarity for similar combinations of the explanatory variables. The use of prior information is central to the Bayesian approach to statistical inference.

We derive estimates of the response surface and measures of precision for the estimates. It is shown how both of these depend critically on the bias term. Two examples illustrate the technique and afford a comparison between the conclusions of the Bayesian model and those of standard response surface models.

The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the author of this report.

BAYESIAN MODELS FOR RESPONSE SURFACES II: ESTIMATING THE RESPONSE SURFACE*

David M. Steinberg

1. INTRODUCTION

We consider the common situation in which an empirical model is sought to describe the relationship between a response variable Y and a collection $X_1,\ldots,X_k$ of explanatory variables. We assume that the true dependence of Y on $\mathbf{x} = (X_1,\ldots,X_k)$ is given by an unknown response function $g(\mathbf{x})$ and that our goal is to obtain a reasonable approximation to g on the basis of experimental data $\{Y_i,\mathbf{x}_i\}_{i=1}^n$. A standard approach to the empirical modeling problem described above is that of response surface methodology (see Box and Wilson 1951, Box 1954, Box and Youle 1955, and Myers 1976), in which a simple graduating function, such as a polynomial of low degree, is used to approximate g. We will consider a generalization of classical response surface models that includes not only a simple graduating function, but also a characterization of the extent to which the graduating function is believed to accurately represent the true response function. The model is Bayesian in that the characterization is expressed in terms of prior distributions.

The model we will analyze is a straightforward extension of a Bayesian model proposed by Blight and Ott (1975) for the special case of polynomial regression with a single explanatory variable. Smith (1973), Young (1977), and O'Hagan (1978) also proposed Bayesian models for estimating an unknown

---

*This work forms part of the author's Ph.D. dissertation written under the direction of Professor G. E. P. Box. The author is grateful to Professor Box for his many valuable comments. The author would also like to thank Professor R. Myers for permission to use the data cited in Section 8.

This research was sponsored by the United States Army under Contract No. DAAG29-80-C-0041. This material is based upon work supported by the National Science Foundation under Grant No. MCS-8210950.

response function. Steinberg (1984) showed that these models are equivalent to the extended Blight-Ott model and that, for certain prior specifications, the models are equivalent to those described by Wahba (1978), which yield generalized smoothing splines as estimates of g. Some of the results in this paper are included in each of the preceding papers. By combining their separate results and some new results, we are able to present a more complete description of Bayesian response surface estimates.

Section 2 describes the model and Section 3 presents the Bayes estimator of the response function g when all the prior distributions in the model are proper. Section 4 studies the implications of assigning improper priors to some of the parameters in the model. Section 5 presents some additional results regarding the vector of estimated values at the design points and Section 6 gives results on the precision of the estimates. Section 7 describes some methods that might be used to estimate additional parameters that appear in the model. Two examples are presented in Section 8 and discussion and concluding comments are offered in Section 9.

2. THE BAYESIAN RESPONSE SURFACE MODEL

Classical response surface models for a response variable Y and explanatory variables $x = (X_1, \ldots, X_k)$ can be written as multiple regression models:

$$Y_x = f'(x)\beta + \varepsilon, \qquad (2.1)$$

where $f(x)$ is a column vector of functions of the explanatory variables, β is a column vector of unknown parameters that must be estimated, ε denotes experimental error, and primes denote vector

(or matrix) transposes. We include the subscript $\mathbf{x}$ on Y in (2.1) to emphasize that this is the model for the response variable at the particular collection $\mathbf{x}$ of explanatory variables. We also write (2.1) as an approximate equality ($\approx$) rather than an exact equality to emphasize that $\mathbf{f}'(\mathbf{x})\boldsymbol{\beta}$ is assumed to be a reasonable local approximation to the true response function $g(\mathbf{x})$, but is not an exact representation. Throughout the paper, we will indicate vectors and matrices with boldface type.

Blight and Ott (1975) suggested that (2.1) could be improved, in the case of polynomial regression, by including an additional term to account for the (in)adequacy of the graduating function $\mathbf{f}'(\mathbf{x})\boldsymbol{\beta}$ to accurately represent $g(\mathbf{x})$. To extend their approach to the general empirical modeling situation described in Section 1 is perfectly straightforward and we do so here. We replace (2.1) by:

$$Y_{\mathbf{x}} = \mathbf{f}'(\mathbf{x})\boldsymbol{\beta} + \eta_{\mathbf{x}} + \varepsilon, \tag{2.2}$$

where $\eta_{\mathbf{x}} = g(\mathbf{x}) - \mathbf{f}'(\mathbf{x})\boldsymbol{\beta}$ is the bias at $\mathbf{x}$ associated with the particular graduating function that has been chosen. We write (2.2), unlike (2.1), with an equals ($=$) sign to emphasize that inclusion of the additional term $\eta_{\mathbf{x}}$ is assumed to make an exact representation of $g(\mathbf{x})$ possible.

We now make the following assumptions about the terms in (2.2):

$$\varepsilon \sim N(0,\sigma^2) \quad \text{i.i.d.} \tag{2.3a}$$

$$\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \mathbf{V}) \tag{2.3b}$$

$$\eta_{\mathbf{x}} \text{ is a continuous Gaussian stochastic process} \tag{2.3c}$$

$$\eta_{\mathbf{x}} \sim N(0, \tau\sigma^2 R(\mathbf{x},\mathbf{x})) \tag{2.3d}$$

$$\mathrm{Cov}(\eta_{\mathbf{u}}, \eta_{\mathbf{v}}) = \tau\sigma^2 R(\mathbf{u}, \mathbf{v}). \qquad (2.3e)$$

Assumption (2.3a) is the common assumption that the random error terms are i.i.d. normal deviates and assumption (2.3b) states that the parameters in the graduating function are assumed, a priori, to have a multivariate normal distribution. Thus assumptions (2.3a,b) are the standard assumptions for a Bayesian multiple regression analysis. A special case that has been afforded special attention (see Lindley and Smith 1972, Blight and Ott 1975, Steinberg 1984) is when the regression coefficients are assigned a vague prior distribution. Following Lindley and Smith (1972), a vague prior can be obtained from (2.3b) by considering limiting forms as $V^{-1} \to 0$.

The special feature of (2.2) is the introduction of the "bias" function $\eta_{\mathbf{x}}$ directly into the statistical model. Assumptions (2.3c-e) are designed to provide a characterization of the extent of the bias and justification for these assumptions is made by a direct appeal to prior belief (see, for example, Blight and Ott 1975, p. 80). First, it is assumed that a graduating function has been chosen which captures the primary features of the dependence of Y on $\mathbf{x}$; therefore, it is reasonable to anticipate that the bias at any given point will be 0. The prior variances in (2.3d) express the believed extent of bias throughout the explanatory variable space. Thus, for example, if the graduating function is thought to provide a good approximation in one region of the factor space, but may be increasingly susceptible to bias outside that region (see, for example, Box and Draper 1959), the prior variances in (2.3d) can

be defined to reflect that belief. Finally, the prior covariances in (2.3e) can be chosen to reflect the likely similarity of the bias at proximate locations in the factor space, which is closely related to prior beliefs about the "smoothness" of the true response function. The particular parameterization for the covariance function emphasizes two aspects: (i) it is natural to measure bias, not absoutely, but relative to the magnitude of experimental error, so σ^2 has been factored out of (2.3e); and (ii) it is natural to further decompose the covariance function into a "standardized" covariance function $R(\mathbf{u},\mathbf{v})$ and a proportionality constant τ that indicates the overall magnitude of the bias, relative to that of experimental error. The standardization of the covariance function might be accomplished in a number of ways. If, for example, $\mathrm{Var}\{\eta_{\mathbf{x}}\}$ is assumed to be constant (as in Smith 1973 and Blight and Ott 1975), then it would be natural to let $R(\mathbf{u},\mathbf{v})$ be the corresponding correlation function. Alternatively, $R(\mathbf{x},\mathbf{x})$ might be set equal to 1 for some particular point $\mathbf{x}$; then τ would measure the relative extent of bias at that point. Assumptions (2.3c-e) generalize those made by Blight and Ott (1975), who proposed a particular parametric form for the covariance function which they felt was appropriate for polynomial regression.

Steinberg (1984) showed how (2.2)-(2.3) is related to models that were proposed by Smith (1973) and O'Hagan (1978) and to the generalized smoothing spline estimates of Wahba (1978), and also showed how (2.2)-(2.3) can be expressed in terms of a generalized

Fourier series for the response function. We briefly indicate the form of these models so that we can relate the results of later sections to them. See Steinberg (1984) for further details.

Smith (1973) proposed a hierarchical model for the nx1 data vector $\mathbf{Y}$:

$$\mathbf{Y}/\boldsymbol{\theta} \sim N(\boldsymbol{\theta}, \sigma^2\mathbf{I}) \tag{2.4a}$$

$$\boldsymbol{\theta}/\boldsymbol{\beta} \sim N(\mathbf{X}\boldsymbol{\beta}, \tau\sigma^2\mathbf{R}) \tag{2.4b}$$

$$\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \mathbf{V}), \tag{2.4c}$$

where $\boldsymbol{\theta}$ is the vector of expected values of the observations, $\mathbf{X}$ is the nxp matrix whose ith row is given by $\mathbf{f}'(\mathbf{x}_i)$, and $\mathbf{R}$ is the nxn matrix whose i,jth entry is $R(\mathbf{x}_i, \mathbf{x}_j)$. O'Hagan (1978) generalized (2.1) by allowing the parameter vector $\boldsymbol{\beta}$ to be a function of $\mathbf{x}$ and used prior distributional assumptions to express the believed variability of $\boldsymbol{\beta}(\mathbf{x})$.

Wahba (1978) proposed estimating $g(\mathbf{x})$ by finding that function $g_{n,\tau}(\mathbf{x})$ in a Hilbert space that minimizes:

$$\sum_{i=1}^{n} \left[Y_i - h(\mathbf{x}_i)\right]^2 + \tau^{-1}J(h), \tag{2.5}$$

where J is a semi-norm on the Hilbert space that is typically a roughness penalty. Wahba proved that the estimate $g_{n,\tau}$ obtained from (2.5) is also the solution to a Bayesian estimation problem and Steinberg (1984) showed that the appropriate Bayesian model is precisely (2.3) when the regression coefficients are assigned a vague prior distribution. (Note that τ in (2.5) is equal to

$(n\lambda)^{-1}$ in Wahba's formulation.) Wecker and Ansley (1983) also studied this model.

Steinberg (1984) showed that (2.3) also results from consideration of a Bayesian model for a generalized Fourier series expansion of $g(x)$:

$$g(x) = \sum_{j=1}^{p} \beta_j f_j(x) + \sum_{i=0}^{\infty} \theta_i g_i(x), \qquad (2.6a)$$

where $\beta \sim N(\beta_0, V)$ (2.6b)

and $\theta_i \sim N(0, \tau\sigma^2 m_i^2)$, independent. (2.6c)

Defining η_x to be the second summation in (2.6a) yields (2.3) with

$$R(u,v) = \sum_{i=0}^{\infty} m_i^2 \, g_i(u) \, g_i(v),$$

provided the above series converges whenever $u = v$. Moreover, (2.3) admits such an expansion whenever the factor space is a compact set.

3. BAYESIAN ESTIMATES WITH PROPER PRIORS

In this section we derive estimates of the response function $g(\mathbf{x})$ and the regression coefficients based on the Bayesian model defined by (2.2)-(2.3) of Section 2. Throughout this section, we will assume that the regression coefficients have been assigned a proper prior distribution; estimation with an improper prior will be discussed in Section 4.

Consider first the problem of estimating the response function g at the point $\mathbf{x}=(X_1,\ldots,X_k)$ in the factor space. The Bayesian model presented in Section 2 represents $g(\mathbf{x})$ as $\mathbf{f}'(\mathbf{x})\boldsymbol{\beta} + \eta_{\mathbf{x}}$ and so provides a prior distribution for $g(\mathbf{x})$. A natural way to estimate $g(\mathbf{x})$, then, is to find its *posterior* distribution; i.e., to find the conditional distribution of $g(\mathbf{x})$ given the observed data. Similarly, estimates of the vector of regression coefficients should be based on the posterior distribution of $\boldsymbol{\beta}$. To find these distributions, denote the experimental responses by the $nx1$ random vector $\mathbf{Y}$. Then:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta} + \boldsymbol{\varepsilon} \tag{3.1a}$$

$$g(\mathbf{x}) = \mathbf{f}'(\mathbf{x})\boldsymbol{\beta} + \eta_{\mathbf{x}} \tag{3.1b}$$

where $\mathbf{X}$ is the nxp matrix whose ith row is $\mathbf{f}'(\mathbf{x}_i)$, $\boldsymbol{\eta} = (\eta_{\mathbf{x}_1},\ldots,\eta_{\mathbf{x}_n})'$, and $\boldsymbol{\varepsilon}$ is the $nx1$ vector of experimental errors for the observed data. Applying (2.3), simple computations reveal that the joint distribution of $(\mathbf{Y}',g(\mathbf{x}),\boldsymbol{\beta}')'$ is multivariate normal with expected value: $\left((\mathbf{X}\boldsymbol{\beta}_0)',\mathbf{f}'(\mathbf{x})\boldsymbol{\beta}_0,\boldsymbol{\beta}_0'\right)'$ and covariance matrix:

$$\begin{array}{c|c|c} \mathbf{XVX'} + \tau\sigma^2\mathbf{R} + \sigma^2\mathbf{I} & \mathbf{XVf(x)} + \tau\sigma^2\mathbf{r(x)} & \mathbf{XV} \\ \hline \mathbf{f'(x)VX'} + \tau\sigma^2\mathbf{r'(x)} & \mathbf{f'(x)Vf(x)} + \tau\sigma^2\mathbf{R(x,x)} & \mathbf{f'(x)V} \\ \hline \mathbf{VX'} & \mathbf{Vf(x)} & \mathbf{V} \end{array}$$

where $\mathbf{R}$ is the nxn matrix whose i,jth entry is $R(\mathbf{x}_i,\mathbf{x}_j)$, $\mathbf{I}$ is the nxn identity matrix, and $\mathbf{r(x)}$ is an nx1 vector whose ith entry is $R(\mathbf{x}_i,\mathbf{x})$.

Since the joint distribution of $(\mathbf{Y}',g(\mathbf{x}),\boldsymbol{\beta}')$ is multivariate normal, the posterior distributions of $g(\mathbf{x})$ and $\boldsymbol{\beta}$ will also be normal. Thus the natural point estimates of $g(\mathbf{x})$ and $\boldsymbol{\beta}$ based on their posterior distributions will be their posterior expectations. Certainly, if the estimation problem is placed in a decision-theoretic framework, the posterior expectations will be the best estimates with respect to all symmetric loss functions. We give the posterior expectations in the following theorem.

Theorem 3.1: Suppose that the model is specified by (2.2)-(2.3) and that the observed data are $\mathbf{Y}=\mathbf{y}$. Denoting the posterior expectation of $g(\mathbf{x})$ by $\hat{g}(\mathbf{x})$, we have:

$$\begin{aligned} \hat{g}(\mathbf{x}) &= E\{g(\mathbf{x})/\mathbf{Y}=\mathbf{y}\} \\ &= \mathbf{f'(x)}\boldsymbol{\beta}_0 + \left[\tau\sigma^2\mathbf{r'(x)} + \mathbf{f'(x)VX'}\right] \\ &\qquad \times (\sigma^2\mathbf{I} + \tau\sigma^2\mathbf{R} + \mathbf{XVX'})^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0). \end{aligned} \tag{3.2}$$

The posterior expectation of $\boldsymbol{\beta}$ is:

$$E\{\boldsymbol{\beta}/\mathbf{Y}=\mathbf{y}\} = \boldsymbol{\beta}_0 + \mathbf{VX'}(\sigma^2\mathbf{I} + \tau\sigma^2\mathbf{R} + \mathbf{XVX'})^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0). \tag{3.3}$$

Proof: The proof follows directly from standard results for conditional expectations of multivariate normal distributions (see, for example, Anderson 1958, p. 28).

Although Theorem 3.1 is straightforward, it is not particularly revealing. The following results provide more intuition into the form of the estimates. We begin by observing how (3.2) and (3.3) are related.

Corollary: (i) The posterior expectation of $g(x)$ has a natural decomposition as the sum of a graduating function whose coefficients are estimated by (3.3) and a second term, which Blight and Ott call the correction for bias, that depends on $r(x)$, τ, and σ^2:

$$\hat{g}(x) = f'(x)E\{\beta/Y=y\} + E\{\eta_x/Y=y\}, \qquad (3.4)$$

where

$$E\{\eta_x/Y=y\} = \tau\sigma^2 r'(x)(\sigma^2 I + \tau\sigma^2 R + XVX')^{-1}(y - X\beta_0) \qquad (3.5)$$

(ii) The bias term (3.5) can be expressed in terms of the residuals at the design points when the graduating function alone is fit. Denote by $\hat{Y}_{GF}$ the vector of predictions that results from fitting just the graduating function: $\hat{Y}_{GF} = XE\{\beta/Y=y\}$, and denote by $\hat{e}_{GF}$ the corresponding residual vector: $\hat{e}_{GF} = y - \hat{Y}_{GF}$. Then

$$E\{\eta_x/Y=y\} = \tau r'(x)(I + \tau R)^{-1}\hat{e}_{GF}.$$

Proof: Part (ii) of the corollary is trivial. Part (i) can be verified directly, using the same approach as in Theorem 3.1. Alternatively, since $g(x) = f'(x)\beta + \eta_x$, it follows that

$$\begin{aligned} E\{g(x)/Y=y\} &= E\{f'(x)\beta/Y=y\} + E\{\eta_x/Y=y\} \\ &= f'(x)E\{\beta/Y=y\} + E\{\eta_x/Y=y\} \end{aligned}$$

which proves (3.4) and, along with equations (3.2) and (3.3), implies (3.5).

The corollary provides valuable insight into the role of the

bias covariance function $\tau\sigma^2 R(u,v)$ in estimating $g(x)$. Substituting this expression into the second term of (3.4) yields:

$$\hat{g}(x) = f'(x)E\{\beta/Y=y\} + \tau\sigma^2 \sum_{i=1}^{n} a_i R(x,x_i), \qquad (3.6)$$

where the coefficients a_i are estimated from the data. Thus the estimation equation, as a function of x, combines a graduating function with a linear combination of n functions that are completely determined by the form of the covariance function $R(u,v)$ and the choice of design points.

Equation (3.6) provides a broad, flexible class of estimators for response surfaces. Moreover, it may suggest useful guidelines for choosing a prior covariance function, since some choices will lead to especially appealing estimation equations while others may have undesirable consequences. For example, Blight and Ott (1975) considered the special case of univariate polynomial regression and suggested using $R(u,v) = \lambda^{|u-v|}$, $0<\lambda<1$, which is the covariance function for a first order autoregressive process. It can be seen from (3.6) that this choice leads to an estimate $\hat{g}(x)$ whose derivative is discontinuous at each design point. If it is believed that the response function has a continuous derivative, then this prior covariance function leads to a poor representation of posterior belief.

Smith's (1973) prior specification for univariate regression, whereby $R = I$, is also called into question. (Smith made prior

assumptions only about the matrix R, not about the complete covariance function.) In general, R will be equal to the identity only if $R(u,v)$ decreases rapidly as $|u - v|$ increases, in which case $R(x,x_i)$ will have a rather sharp "spike" around x_i. From (3.6), the resulting estimate $\hat{g}(x)$ will deviate from the graduating function only in the vicinity of the design points, but these deviations may be quite sharp. This also would seem to be an unlikely summary of posterior belief about the nature of the response function.

O'Hagan (1978) also considered in detail the case of univariate polynomial regression and suggested a covariance function of the form:

$$R(u,v) = \exp\{-(u-v)^2/d^2\} \times f'(u)Wf(v),$$

where d^2 is a scaling parameter and W is a positive definite $p \times p$ matrix. This choice is similar to Blight and Ott's, but there are two differences. First, $R(u,u)$ will now typically be an increasing function of u^{2k}, where k is the degree of the polynomial, instead of a constant, as in Blight and Ott's specification; second, $R(u,v)$ now decreases exponentially in the square of $|u - v|$. An important consequence of the latter change is that $\hat{g}(x)$ will now be an analytic function of x instead of a function with a discontinuous first derivative.

Wahba (1978) devoted particular attention to univariate polynomial regression on the interval $(0,1)$ and recommended choosing R to produce polynomial spline estimates. In particular,

for the popular cubic splines, the graduating function is a straight line and:

$$R(u,v) = \begin{cases} (3u^2v - u^3)/6 & \text{if } u \leq v \\ (3uv^2 - v^3)/6 & \text{if } v \leq u, \end{cases}$$

which is the covariance function for an integrated Brownian motion on the unit interval (see Shepp 1966). Note that for fixed v, $R(u,v)$ is itself a cubic spline as a function of u; since a linear combination of cubic splines is again a cubic spline, it follows from (3.6) that this choice of $R(u,v)$ will lead to an estimate $\hat{g}(x)$ which is a cubic spline. (Note, however, that (3.6) will <u>not</u> give Wahba's cubic spline estimate of $g(x)$, which results only when the regression coefficients are assigned a vague prior; the results above show that even with a proper prior, it is possible to obtain an estimate that is a cubic spline.)

If a Bayesian generalized Fourier series representation of $g(x)$ is used as a basis to derive the covariance function, then the estimate $\hat{g}(x)$ can also be written as a generalized Fourier series. In particular, if the experiment is modeled by (2.6a-c), then substituting (2.7) into (3.6) yields:

$$\hat{g}(x) = f'(x)E\{\beta/\mathbf{Y}=\mathbf{y}\} + \sum_{j=0}^{\infty} b_j g_j(x), \qquad (3.7)$$

where $b_j = \tau\sigma^2 \sum_{i=1}^{n} a_i g_j(x_i)$. The $\{a_i\}$ in the last expression are the same coefficients that appear in (3.6). Moreover, it is easy to

show that:

$$b_j = E\{\theta_j / \mathbf{Y}=\mathbf{y}\}.$$

Thus, if the generalized Fourier series approach is employed, $\hat{g}(\mathbf{x})$ is a series of exactly the same form as $g(\mathbf{x})$, but with the coefficients of all the terms in the series estimated by their posterior means.

4. BAYESIAN ESTIMATES WITH IMPROPER PRIORS

The authors of the Bayesian models that were described in Section 2 have argued that it is often appropriate to assume an improper prior distribution for the regression coefficients β. In this section, we explore the consequences of such an assumption for estimating the response surface and the regression parameters. For discussion of the significance of assigning β an improper prior, see Steinberg (1984, Sect. 5.3).

As noted in Section 2, a natural way to assign β an improper prior is to consider limiting forms as its prior precision matrix $\mathbf{V}^{-1}$ converges to a $\mathbf{0}$ matrix; i.e., to retain the form of a normal distribution but to allow the prior variance of that distribution to become arbitrarily large. We can then derive the posterior distributions of $g(\mathbf{x})$ and β by taking limits of the corresponding posterior distributions when a proper prior is used. Specifically, we will derive limiting values for their posterior means (in this section) and variances (in Section 6) and will show that these limits exist provided that the $\mathbf{X}$ matrix has full column

rank. Since the posterior distributions are normal when a proper prior is used, the corresponding posterior densities must converge to a normal density with the limiting mean and variance as its parameters. Applying a theorem of Scheffé (1947), we conclude that the limiting posterior distributions are also normal. As such, the posterior means will be the natural estimates of the response function and the regression parameters and we give these in the following theorem.

Theorem 4.1: Suppose X has full column rank and an improper prior is assigned to the regression coefficients. Then the posterior mean of $g(x)$ is:

$$\hat{g}(x) = \lim_{V^{-1} \to 0} E\{g(x)/Y=y\} \tag{4.1}$$
$$= f'(x)\left[X'M^{-1}X\right]^{-1}X'M^{-1}y +$$
$$\tau\sigma^2 r'(x)\{M^{-1} - M^{-1}X\left[X'M^{-1}X\right]^{-1}X'M^{-1}\}y,$$

where $M = (I + \tau R)$. The estimated regression coefficients are:

$$\lim_{V^{-1} \to 0} E\{\beta/Y=y\} = \left[X'M^{-1}X\right]^{-1}X'M^{-1}y. \tag{4.2}$$

Before proceeding with the proof, we note that this result is by no means new; in fact, Theorem 4.1 is only a slight generalization of Theorem 2 in Wahba (1978) and Theorem 2 in O'Hagan (1978), both of which assume that $V = kI$.

Proof: The proof relies on two simple matrix identities, which we state as lemmas. In both of the lemmas, we assume that the dimensions of the matrices are such that the indicated matrix operations are well-defined and that all the necessary matrix inverses exist.

Lemma 1: $(A + XVX')^{-1} = A^{-1} - A^{-1}X(X'A^{-1}X + V^{-1})^{-1}X'A^{-1}$. (4.3)

Lemma 2: $VX'(A + XVX')^{-1} = (X'A^{-1}X + V^{-1})^{-1}X'A^{-1}$. (4.4)

Both of the lemmas can be verified directly; an interesting statistical proof of Lemma 1 was given by Lindley and Smith (1972) using properties of a hierarchical Bayesian linear model. The proof of the main result follows directly upon using the lemmas to rewrite (3.2) and (3.3) with $A = \sigma^2 M = \sigma^2(I + \tau R)$, noting that:

$$(X'A^{-1}X + V^{-1})^{-1} \to \sigma^2(X'M^{-1}X)^{-1} \quad \text{as} \quad V^{-1} \to 0$$

provided that X has full column rank. Details are given in the Appendix. Whereas earlier proofs required several complicated limits, only the simple limiting result above is needed here.

As with the estimates (3.1) and (3.2) (when proper priors are used), (4.1) can be decomposed into the sum of a graduating function whose coefficients are estimated by (4.2) and a "correction for bias" term:

$$\lim_{V^{-1} \to 0} E\{g(x)/Y=y\} = f'(x) \lim_{V^{-1} \to 0} E\{\beta/Y=y\} + \lim_{V^{-1} \to 0} E\{\eta_x/Y=y\}, \tag{4.5}$$

where the latter term is given by the final line of (4.1). Also, the estimated response function can once again be written, as in (3.6), as the sum of a graduating function and n functions whose form is completely determined by the covariance function (2.3e):

$$\lim_{V^{-1} \to 0} E\{g(x)/Y=y\} = f'(x) \lim_{V^{-1} \to 0} E\{\beta/Y=y\} + \tau \sum_{i=1}^{n} a_i R(x,x_i), \tag{4.6}$$

where the coefficients a_i are estimated from the data. Thus all the comments of the previous section on the role of the covariance function $R(u,v)$ in the estimated response function are also valid if the regression coefficients are assigned an improper prior. In particular, if the covariance function is derived using a generalized Fourier series model, the estimate will be a generalized Fourier series of the same form whose coefficients will be estimated by: $\lim_{V^{-1} \to 0} E\{\theta_j/Y=y\}$.

There are two important differences in the estimated response function that result from assuming a vague prior distribution for the regression coefficients. First, the estimation equation no longer depends on the prior mean β_0 for these coefficients. The regression coefficients are now estimated on the basis of the data alone. Second, the estimation equation is independent of σ^2, but does depend on the parameter τ which expresses the overall magnitude of bias relative to that of experimental error.

The estimate (4.2) for the regression coefficients is clearly the generalized least squares estimate that corresponds to the model $Y = X\beta + \epsilon$ in which the error terms are correlated with $E\{\epsilon\epsilon'\} = \sigma^2(I + \tau R)$, and is the maximum likelihood estimate for this model if it is further assumed that the errors have a multivariate normal distribution. These facts suggest a

justification for (4.2) as an estimator of β in classical sampling theory terms. If the proposed regression model is only an approximation to the true response function, then deviations from that model will be the sum of two components, one due to experimental error and the other due to bias. It is common to assume that experimental errors are independent of one another, but such an assumption seems quite implausible for errors due to bias. Thus the resulting model should involve correlated error terms in which the correlations reflect the extent of the bias, just as in the above model. Of course, such an argument does not suggest what the precise form of the error covariance matrix should be. (One possibility might be to assume a simple parametric form for the covariance matrix and then to maximize the resulting likelihood function over all the parameters.) Also, the frequentist approach would lead to an estimator that includes only the first term of (4.1), the best fitting graduating function in light of experimental error and bias; the second term, which gives the adjustment for bias, arises only in the Bayesian context.

Allowing τ to range from 0 to infinity permits us to model a wide range of situations, from those in which experimental error is dominant (when, say, scientific knowledge provides an exact form for the response function) through those in which the bias is dominant (as might be the case in numerical analysis). We now consider the estimates that result in these limiting cases.

Theorem 4.2: Given the model (2.2)-(2.3) with an improper prior

assigned to the regression coefficients, the estimates (4.1) and (4.2) have the following limiting forms as $\tau \to 0$:

$$\lim_{\tau \to 0} \lim_{V^{-1} \to 0} E\{g(x)/Y=y\} = f'(x)(X'X)^{-1}X'y. \tag{4.7}$$

$$\lim_{\tau \to 0} \lim_{V^{-1} \to 0} E\{\beta/Y=y\} = (X'X)^{-1}X'y. \tag{4.8}$$

If R is non-singular, then the estimates (4.1) and (4.2) have the following limiting forms as $\tau \to \infty$:

$$\lim_{\tau \to \infty} \lim_{V^{-1} \to 0} E\{g(x)/Y=y\} = f'(x)(X'R^{-1}X)^{-1}X'R^{-1}y + r'(x)\{R^{-1} - R^{-1}X(X'R^{-1}X)^{-1}X'R^{-1}\}y. \tag{4.9}$$

$$\lim_{\tau \to \infty} \lim_{V^{-1} \to 0} E\{\beta/Y=y\} = (X'R^{-1}X)^{-1}X'R^{-1}y. \tag{4.10}$$

Proof: Equations (4.7) and (4.8) follow directly from (4.1) and (4.2) upon noting that $(I + \tau R)^{-1} \to I$ as $\tau \to 0$. To obtain equations (4.9) and (4.10), we rewrite (4.1) and (4.2), replacing $(I + \tau R)^{-1}$ by $\tau^{-1}(\tau^{-1}I + R)^{-1}$. After some cancellation, the expressions now depend on τ only through $(\tau^{-1}I + R)^{-1}$, which converges to R^{-1} as $\tau \to \infty$, provided that R is non-singular, and results in (4.9) and (4.10).

The first half of Theorem 4.2 yields familiar answers: the ordinary least squares estimators. Thus, ordinary least squares obtains as a limiting case of the Bayesian model when the regression parameters have an improper prior and when the bias is assumed to be negligible relative to experimental error (i.e. when the graduating function is assumed to exactly represent the response function). The second half of Theorem 4.2 does not have so immediate an

interpretation, but the following section will clarify what happens when τ tends to infinity.

5. SPECIAL FORMS FOR THE $\hat{Y}$ VECTOR

In this section we derive some results describing the vector $\hat{Y}$ of precicted values whose ith entry is $\hat{g}(x_i)$, where x_i is the ith design point. We begin by examining the relationship between $\hat{Y}$ and τ, the presumed extent of bias relative to experimental error, when the regression coefficients are assigned an improper prior.

Theorem 5.1: Suppose, given the model (2.2)-(2.3), that X has full column rank and that $R = (R(x_i,x_j))$ is non-singular. Then:

$$\lim_{\tau\to\infty} \lim_{V^{-1} \to 0} \hat{Y} = y; \tag{5.1}$$

i.e., the estimation equation interpolates the observed data. Replicate observations at any design points will contribute identical rows and columns to R, making it singular. Suppose, however, that elimination of all identical rows and columns yields a non-singular matrix. Then:

$$\hat{Y}_i = \text{average of all observations at } x_i. \tag{5.2}$$

Proof: Noting that the ith row of X is $f'(x_i)$, and that the ith row of R is $r'(x_i)$, (5.1) is an immediate consequence of (4.9). To obtain (5.2), observe that if there are replicate observations at x_i, all the information they provide about $g(x)$ is contained in their average. Thus we can compute (3.2), (4.1),

and (4.9) conditioning only on the vector of replicate averages, rather than the entire data vector, which leads to (5.2). Details are given in the Appendix.

Combining the results of Theorem 4.2 and Theorem 5.1 provides considerable insight into how the estimated response function $\hat{g}(\mathbf{x})$ behaves as a function of τ: it varies from the least squares graduating function (when $\tau = 0$) to an interpolant (when $\tau \to \infty$). As an intuitive justification for the latter result, we might think of $\tau \to \infty$ as an appropriate way to model data that are *not* subject to experimental error, so that the observed responses are exact values of the response function. The estimated response function reflects this certain knowledge by correctly estimating the response at those points.

The above description of how the estimated response function depends on τ is well-known in the spline literature when all the design points are distinct (see, for example, Kimeldorf and Wahba 1971), but has generally not been extended to cover replicates nor has it been observed in connection to the Bayesian models. Blight and Ott, for example, proposed a parametric form for the covariance function R and then suggested that these parameters and τ be jointly estimated by minimizing the residual sum of squares, $S(\mathbf{R},\tau) = \sum (y_i - \hat{Y}_i)^2$. It is clear from Theorem 5.1 that $S(\mathbf{R},\tau)$ will always be minimized when $\tau \to \infty$, regardless of the values of the other parameters.

A common property of Bayes estimates using proper priors is

that they can be expressed as weighted averages of their prior means and the observed data. This is not possible for $\hat{g}(\mathbf{x})$ for an arbitrary point $\mathbf{x}$ because, in general, no observation has been made there; for the n observed data points, however, and for the estimated regression coefficients, we can write such weighted averages.

Theorem 5.2: Given the model (2.2)-(2.3), the predicted value vector $\hat{\mathbf{Y}}$ can be expressed as a weighted average of its prior mean $\mathbf{X}\boldsymbol{\beta}_0$ and the observed response vector $\mathbf{y}$, where the weights are inversely proportional to the respective measures of variation, $\tau\sigma^2\mathbf{R} + \mathbf{X}\mathbf{V}\mathbf{X}'$ and σ^2:

$$\hat{\mathbf{Y}} = \left[\sigma^{-2}\mathbf{I} + (\tau\sigma^2\mathbf{R} + \mathbf{X}\mathbf{V}\mathbf{X}')^{-1}\right]^{-1} \times \left[\sigma^{-2}\mathbf{y} + (\tau\sigma^2\mathbf{R} + \mathbf{X}\mathbf{V}\mathbf{X}')^{-1}\mathbf{X}\boldsymbol{\beta}_0\right]. \qquad (5.3)$$

Similarly, the estimated regression coefficients can be written as a weighted average of their prior mean $\boldsymbol{\beta}_0$ and the observed responses, with the weights inversely proportional to the prior and data covariance matrices, respectively:

$$E\{\boldsymbol{\beta}/\mathbf{Y}=\mathbf{y}\} = \left[\sigma^{-2}\mathbf{X}'(\mathbf{I} + \tau\mathbf{R})^{-1}\mathbf{X} + \mathbf{V}^{-1}\right]^{-1} \times \left[\sigma^{-2}\mathbf{X}'(\mathbf{I} + \tau\mathbf{R})^{-1}\mathbf{y} + \mathbf{V}^{-1}\boldsymbol{\beta}_0\right]. \qquad (5.4)$$

Proof: We exploit here the hierarchical model formulation (2.4a-c) proposed by Smith (1973). In this formulation, $E\{\mathbf{Y}/\boldsymbol{\theta}\} = \boldsymbol{\theta}$, so that $\hat{\mathbf{Y}} = E\{\boldsymbol{\theta}/\mathbf{Y}=\mathbf{y}\}$. Equation (5.3) is then a special case of a theorem proved by Lindley and Smith (1972, equations 12 and 13). To prove (5.4), we combine (2.4a) and (2.4b) to obtain the distribution of $\mathbf{Y}$ conditional on $\boldsymbol{\beta}$ without the mediating parameter $\boldsymbol{\theta}$:

$$Y/\beta \sim N(X\beta, \sigma^2(I + \tau R)).$$

This, together with (2.4c) matches the assumptions of a lemma proved by Lindley and Smith (1972); their equations (7) and (8) imply (5.4).

Steinberg (1984), exploiting results of Wahba (1978), showed that the Bayesian model considered here leads to generalized spline estimates when the regression coefficients are assigned an improper prior. The underlying motivation for spline estimates is to find a reasonably "smooth" function that closely follows the observed data. Spline estimates are derived as solutions to the minimization problem (2.5) stated in Section 2. In the following theorem, we show that $\hat{Y}$ is the solution to a discrete analogue of (2.5).

Theorem 5.3: $\hat{Y}$ solves the minimization problem: find u to minimize

$$(u-y)'(u-y) + (u-X\beta_0)'\sigma^2(\tau\sigma^2 R + XVX')^{-1}(u-X\beta_0). \quad (5.5)$$

As $V^{-1} \to 0$, (5.5) tends to:

$$(u-y)'(u-y) + \tau^{-1}u'\left[R^{-1} - R^{-1}X(X'R^{-1}X)^{-1}X'R^{-1}\right]u. \quad (5.6)$$

Moreover, the second term in (5.6) is 0 if and only if $u \in \text{col}(X)$; that is, if and only if u can be written as a linear combination of the columns of X.

Proof: The proof is given in the Appendix.

Both (5.5) and (5.6) describe the vector $\hat{Y}$ of predicted values as the solution to a minimization problem composed of two terms: the residual sum of squares, $(\hat{Y}-y)'(\hat{Y}-y)$, and a quadratic penalty term. For the general case (5.5), the quadratic term

penalizes $\hat{\mathbf{Y}}$ in accord with its distance from its prior expectation, $\mathbf{X}\beta_0$. This has the effect of "shrinking" the vector of predicted values toward its prior expectation. The extent of the shrinkage depends on the weighting matrix $\sigma^2(\tau\sigma^2\mathbf{R} + \mathbf{XVX}')^{-1}$, which is proportional to σ^2, but is inversely proportional to the prior variance $\tau\sigma^2\mathbf{R} + \mathbf{XVX}'$. Thus the prior expectation will be most influential when our prior precision is great relative to experimental error; when our prior precision is not great, the data will dominate the prior in determining $\hat{\mathbf{Y}}$.

The quadratic penalty term undergoes several interesting changes in the limiting case of (5.6). First, the penalty depends on the variance-bias tradeoff parameter τ but not on σ^2. Second, the penalty is independent of the prior expectation $\mathbf{X}\beta_0$. Third, the penalty is 0 only for those vectors of predicted values which are in the column space of $\mathbf{X}$; i.e., for those vectors of predicted values which can be exactly written as a graduating function. The meaning of these last two points is that the penalty no longer induces shrinkage toward a particular, pre-specified vector of predicted values; rather, in a more general sense, there is shrinkage toward the response plane spanned by the graduating function. Finally, equation (5.6) is an exact discrete analogue of (2.5), the continuous smoothing problem that leads to generalized spline estimation. Thus Theorem 5.3 further illustrates the close link between spline estimation and the Bayesian models when a diffuse prior is assigned to the regression coefficients.

Theorem 5.3 can also be used to show how the residual sum of squares depends on the choice of τ. Let us denote the vector of predicted values by $\hat{\mathbf{Y}}(\tau)$ to emphasize its dependence on τ. Then define the residual sum of squares function by

$$RSS(\tau) = [\mathbf{y} - \hat{\mathbf{Y}}(\tau)]'[\mathbf{y} - \hat{\mathbf{Y}}(\tau)].$$

Corollary: If $\mathbf{X}$ has full column rank and the regression coefficients are assigned an improper prior, then $RSS(\tau)$ is a monotone decreasing function of τ, with $RSS(0)$ equal to the residual sum of squares from fitting the graduating function by ordinary least squares. If, in addition, $\mathbf{R}$ is non-singular, then $RSS(\infty) = 0$.

6. PRECISION OF THE ESTIMATES

Measures of precision for the Bayesian model described in Section 2 can be obtained in exactly the same fashion as the estimates derived in the preceding sections. Since the posterior distributions of $g(x)$ and of β are normal, the natural measures of precision are the corresponding posterior variances, which we give in the following theorem.

Theorem 6.1 Given the model (2.2)-(2.3), the posterior variance of $g(x)$ is:

$$\begin{aligned} \mathrm{Var}\{g(x)/\mathbf{Y}=\mathbf{y}\} &= \tau\sigma^2 R(x,x) + f'(x)\mathbf{V}f(x) \\ &\quad - [\tau\sigma^2 r'(x) + f'(x)\mathbf{V}\mathbf{X}'][\sigma^2\mathbf{I} + \tau\sigma^2\mathbf{R} + \mathbf{X}\mathbf{V}\mathbf{X}']^{-1} \\ &\quad \times [\tau\sigma^2 r(x) + \mathbf{X}\mathbf{V}f(x)]. \end{aligned} \tag{6.1}$$

The posterior variance matrix for β is:

$$\mathrm{Var}\{\beta/\mathbf{Y}=\mathbf{y}\} = \mathbf{V} - \mathbf{V}\mathbf{X}'(\sigma^2\mathbf{I} + \tau\sigma^2\mathbf{R} + \mathbf{X}\mathbf{V}\mathbf{X}')^{-1}\mathbf{X}\mathbf{V}. \tag{6.2}$$

Proof: The proof, like that of Theorem 3.1, follows from standard properties of multivariate normal vectors.

The posterior variance of $g(x)$, like the posterior mean, can be decomposed into several terms. Following equation (2.2),

$$\begin{aligned} \mathrm{Var}\{g(x)/\mathbf{Y}=\mathbf{y}\} &= \mathrm{Var}\{f'(x)\beta + \eta_x/\mathbf{Y}=\mathbf{y}\} \\ &= f'(x)\mathrm{Var}\{\beta/\mathbf{Y}=\mathbf{y}\}f(x) + \mathrm{Var}\{\eta_x/\mathbf{Y}=\mathbf{y}\} \\ &\quad + 2f'(x)\mathrm{Cov}\{\beta,\eta_x/\mathbf{Y}=\mathbf{y}\}. \end{aligned} \tag{6.3}$$

Once again, it is particularly interesting to study the special case when the regression coefficients are assigned an improper prior. As we noted in Section 4, the posterior distributions are still normal, provided the $\mathbf{X}$ matrix has full column rank. In the

following theorem, we derive the posterior variances.

Theorem 6.2: Suppose $\mathbf{X}$ has full column rank and an improper prior is assigned to the regression coefficients of (2.2)-(2.3). Then the posterior variance of $g(\mathbf{x})$ is:

$$\lim_{\mathbf{V}^{-1} \to 0} \mathrm{Var}\{g(\mathbf{x})/\mathbf{Y}=\mathbf{y}\} = \sigma^2\{\tau R(\mathbf{x},\mathbf{x}) + \mathbf{f}'(\mathbf{x})(\mathbf{X}'\mathbf{M}^{-1}\mathbf{X})^{-1}\mathbf{f}(\mathbf{x}) - 2\tau\mathbf{r}'(\mathbf{x})\mathbf{M}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{M}^{-1}\mathbf{X})^{-1}\mathbf{f}(\mathbf{x}) - \tau^2\mathbf{r}'(\mathbf{x})[\mathbf{M}^{-1} - \mathbf{M}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{M}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}^{-1}]\mathbf{r}(\mathbf{x})\}, \tag{6.4}$$

where $\mathbf{M} = \mathbf{I} + \tau\mathbf{R}$. The posterior variance matrix of $\boldsymbol{\beta}$ is:

$$\lim_{\mathbf{V}^{-1} \to 0} \mathrm{Var}\{\boldsymbol{\beta}/\mathbf{Y}=\mathbf{y}\} = \sigma^2(\mathbf{X}'\mathbf{M}^{-1}\mathbf{X})^{-1}. \tag{6.5}$$

Proof: The proof is given in the Appendix.

When the regression coefficients are assigned an improper prior, we found in Section 4 that the posterior means are independent of σ^2; we see above that the posterior variances are proportional to σ^2 with a constant of proportionality that is a function of τ, the covariance function $R(\mathbf{u},\mathbf{v})$, the experimental design and, in the case of $g(\mathbf{x})$, the point $\mathbf{x}$. The posterior variance matrix for $\boldsymbol{\beta}$ is the same matrix that would result from a classical sampling theory model in which the error terms are correlated with $E\{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\} = \sigma^2(\mathbf{I} + \tau\mathbf{R})$. Although the posterior variances do depend on the experimental design, they are independent of the observed responses. For the special case when $\mathbf{R}$ is non-singular, we note that Wahba (1983, Theorem 2) gives an alternative expression for (6.4). The equivalence of her formula

and (6.4) can be verified by straightforward, but tedious, algebra.

It is quite difficult to describe precisely how the posterior variances behave as functions of the bias covariance parameters, the design, and the estimation site. Some general conclusions, however, can be reached. The following theorem describes how the posterior variances depend on τ.

Theorem 6.3: Assume the data are modeled by (2.2)-(2.3). Then the following conclusions hold:

(i) The posterior variance of $g(\mathbf{x})$ is a monotone increasing function of τ.

(ii) The posterior variance of $g(x)$ obtains a minimum value of

$$\sigma^2 \mathbf{f}'(\mathbf{x})(\mathbf{X}'\mathbf{X} + \sigma^2\mathbf{V}^{-1})^{-1}\mathbf{f}(\mathbf{x})$$

when $\tau = 0$. If the regression coefficients are assigned an improper prior, the minimum value is

$$\sigma^2 \mathbf{f}'(\mathbf{x})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{f}(\mathbf{x}).$$

(iii) If $\mathbf{x}$ is a design point, then:

$$\mathrm{Var}\{\hat{g}(\mathbf{x})/\mathbf{Y}=\mathbf{y}\} \leq \sigma^2.$$

(iv) Suppose the design includes m distinct points, $\mathbf{x}_1,\ldots,\mathbf{x}_m$. Denote by $\bar{\mathbf{Y}}$ the mx1 vector of average responses at the distinct design points; i.e., $\bar{\mathbf{Y}}_i$ = average of observations at $\mathbf{x}_i$. Denote by $\tau\sigma^2\bar{\mathbf{R}}$ the bias covariance matrix of $\bar{\mathbf{Y}}$ and denote by $\tau\sigma^2\bar{\mathbf{R}}_{\mathbf{x}}$ the bias covariance matrix of $(\bar{\mathbf{Y}}', g(\mathbf{x}))$. If both $\bar{\mathbf{R}}$ and $\bar{\mathbf{R}}_{\mathbf{x}}$ are non-singular, then the posterior variance of $g(\mathbf{x})$ diverges to infinity as $\tau \to \infty$.

(v) The posterior variance of β is also a monotone increasing

function of τ; i.e., if we denote (6.2) by $D(\tau)$ to emphasize its dependence on τ, then $D(\tau_2) - D(\tau_1)$ is a positive semi-definite matrix whenever $\tau_2 > \tau_1$.

(vi) The minimum posterior variance of β is attained when $\tau = 0$ and is

$$\sigma^2(\mathbf{X}'\mathbf{X} + \sigma^2\mathbf{V}^{-1})^{-1}.$$

If the regression coefficients are assigned an improper prior, the minimum value is

$$\sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

Proof: The proof is given in the Appendix.

The monotonicity property proved in Theorem 6.3 is intuitively appealing. If we recall that τ reflects the suspected extent of bias relative to experimental error, then increasing τ (all else fixed) corresponds to positing a model in which the effect of bias is more severe. As we might expect, such an assumption leads to a degradation in the precision of the estimates. In particular, Theorem 6.3 allows us to compare models that include a bias term ($\tau > 0$) with models that include only a graduating function ($\tau = 0$). The (tentative) assumption that a particular graduating function exactly describes the response function must lead to a more optimistic assessment of the precision of the estimates than would be obtained if possible bias were explicitly included in the model. In particular, the estimation variances provided by ordinary least squares (OLS) regression analysis are identical to the posterior variances when $\tau = 0$ and the regression coefficients are

assigned an improper prior. Thus OLS leads to a "best case" assessment of precision that may be unduly optimistic.

Another interesting conclusion from Theorem 6.3 concerns the contrasting behavior of the posterior variance of $g(x)$ at design points and at other points. Whereas the former is bounded by σ^2, the latter may diverge to infinity as $\tau \to \infty$. The key term in determining the limiting behavior of $\mathrm{Var}\{g(x)/\mathbf{Y}=\mathbf{y}\}$ is $\mathrm{Var}\{\eta_x/\mathbf{Y}=\mathbf{y}\}$, the posterior variance of the bias term at x, which depends on $R(u,v)$ and the experimental design used. If the posterior variance of the bias is positive, then $\mathrm{Var}\{g(x)/\mathbf{Y}=\mathbf{y}\}$ will diverge to infinity as $\tau \to \infty$. In general, the only way to assure a minimal level of precision for estimating $g(x)$ is to take an observation at x. This contrasts markedly with the implication, when it is assumed that no bias is present, that maximal precision for estimating $g(x)$ can sometimes be obtained by taking observations far away from x.

Useful information about the form of the posterior variances can be obtained by considering the vector of estimated values at the n design points. The basic results are given in the following theorem, which is also found in Wahba (1983) for the special case when $\mathbf{R}$ is non-singular and the regression coefficients are assigned an improper prior.

Theorem 6.4: Define $\boldsymbol{\theta}$ as in (2.4) to be the vector of expected values at the design points. The posterior variance matrix of $\boldsymbol{\theta}$ is:

$$\text{Var}\{\theta/\mathbf{Y}=\mathbf{y}\} = \sigma^2\mathbf{I} - \sigma^4(\sigma^2\mathbf{I} + \tau\sigma^2\mathbf{R} + \mathbf{X}\mathbf{V}\mathbf{X}')^{-1}. \qquad (6.6)$$

(6.6) is monotone increasing in τ (i.e., $\tau_2 > \tau_1$ implies that the difference of the respective variance matrices is positive semi-definite) and achieves a minimum of

$$\sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X} + \sigma^2\mathbf{V}^{-1})^{-1}\mathbf{X}'$$

when $\tau = 0$. If $\mathbf{R}$ is non-singular, (6.6) converges to $\sigma^2\mathbf{I}$ as $\tau \to \infty$. If the regression coefficients are assigned an improper prior distribution, then the limiting posterior covariance matrix is:

$$\lim_{\mathbf{V}^{-1} \to \mathbf{0}} \text{Var}\{\theta/\mathbf{Y}=\mathbf{y}\} = \sigma^2\{\mathbf{I} - \mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{X}[\mathbf{X}'\mathbf{M}^{-1}\mathbf{X}]^{-1}\mathbf{X}'\mathbf{M}^{-1}\}. \qquad (6.7)$$

(6.7) is also monotone increasing in τ and achieves a minimum of $\sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ when $\tau = 0$. If $\mathbf{R}$ is non-singular, (6.7) also converges to $\sigma^2\mathbf{I}$ as $\tau \to \infty$.

Proof: The proof is given in the Appendix.

Several conclusions can be drawn from the limiting forms in Theorem 6.4. If $\tau = 0$, the Bayesian model produces the familiar formulas for Bayesian multiple regression with a proper prior (6.6) or for OLS regression (6.7) when an improper prior is employed. At the other extreme, as $\tau \to \infty$, $\mathbf{R}$ non-singular implies that the posterior variance matrix of θ is $\sigma^2\mathbf{I}$. One possible interpretation of this result is that $\tau \to \infty$ corresponds to a situation in which the experimenter believes that the response function is so erratic that observations at different points are essentially samples from independent, unrelated distributions. Thus

only observations made at $\mathbf{x}$ provide any basis for inference about $g(\mathbf{x})$ and the posterior variance of $g(\mathbf{x})$ at each design point is equal to σ^2, the variance of the observation made there.

7. ESTIMATION OF VARIANCE PARAMETERS

Thus far we have presented results for the Bayesian model of Section 2 that depend on the experimental error variance σ^2, the parameter τ that reflects the magnitude of bias relative to experimental error, and the form of the bias covariance function, $R(\mathbf{u},\mathbf{v})$. In this section we consider various methods that might be used to make inferences about these components of the model. The discussion here will be rather general; we hope to present more exact results in a later paper.

7.1 Estimating σ^2

Ideally we would like to use replicate observations to estimate σ^2. Just as in ordinary least squares regression, replicate observations allow us to form a "pure error" sum of squares that provides an estimate of σ^2 independent of any assumptions about the nature of the response function. Bayes estimates can be obtained by assuming a prior distribution for σ^2 and combining the prior with the information in the replicates.

If there are no replicates in the experimental design, estimates of σ^2 can be based on the vector of residuals $\mathbf{e} = \mathbf{y} - \hat{\mathbf{Y}}$ from the estimated response function. Wahba (1983), in

the context of cubic spline regression, suggested the estimate:

$$\hat{\sigma}^2 = \mathbf{e}'\mathbf{e}/\mathrm{tr}(\mathbf{B}), \tag{7.1}$$

where $\mathbf{B}$ is the matrix that maps $\mathbf{y}$ into $\mathbf{e}$ and tr denotes matrix trace. If $\tau = 0$, (7.1) reduces to the conventional least squares regression estimate of σ^2. The rationale for dividing by $\mathrm{tr}(\mathbf{B})$ is that $\mathrm{Var}\{\mathbf{e}\} = \sigma^2\mathbf{B}$.

Bayes estimates of σ^2 can be derived by assuming an appropriate prior distribution for σ^2. As one possibility, suppose we adopt a uniform prior for σ^2. Then the posterior density of σ^2 under (2.2)-(2.3), up to a proportionality constant, is:

$$p(\sigma^2/\mathbf{Y}) \propto \frac{\exp\{-(\mathbf{Y}-\mathbf{X}\beta_0)'[\sigma^2\mathbf{I} + \tau\sigma^2\mathbf{R} + \mathbf{X}\mathbf{V}\mathbf{X}']^{-1}(\mathbf{Y}-\mathbf{X}\beta_0)/2\}}{\{\det(\sigma^2\mathbf{I} + \tau\sigma^2\mathbf{R} + \mathbf{X}\mathbf{V}\mathbf{X}')\}^{1/2}} \tag{7.2}$$

Since (7.2) results in a rather complicated posterior distribution for σ^2, we might, following Lindley and Smith (1972), estimate σ^2 by the mode of the posterior density. Differentiating (7.2) with respect to σ^2 and equating to 0 yields the equation:

$$(\mathbf{Y}-\mathbf{X}\beta_0)'\mathbf{C}^{-1}(\mathbf{I} + \tau\mathbf{R})\mathbf{C}^{-1}(\mathbf{Y}-\mathbf{X}\beta_0) - \mathrm{tr}[(\mathbf{I} + \tau\mathbf{R})\mathbf{C}^{-1}] = 0, \tag{7.3}$$

where $\mathbf{C} = \sigma^2\mathbf{I} + \tau\sigma^2\mathbf{R} + \mathbf{X}\mathbf{V}\mathbf{X}'$. In general, (7.3) does not provide a closed form solution for σ^2. We can, however, state a closed form for the special case in which the regression coefficients are assigned an improper prior. As in Section 4, we consider the limit of the left hand side of (7.3) as $\mathbf{V}^{-1}$ converges to a $\mathbf{0}$ matrix. Using Lemma 1 of Section 4, it is easily verified that $\mathbf{C}^{-1} \to \sigma^{-2}\mathbf{B}$ as $\mathbf{V}^{-1} \to \mathbf{0}$, where $\mathbf{B}$ is the matrix referred to above

that maps y into the residual vector e. Further, $(I + \tau R)B$ is an idempotent matrix that projects into the orthogonal complement of the column space of X, so that the second term in (7.3) converges to $(n - p)/\sigma^2$. Thus, the posterior modal estimate when the regression coefficients are assigned an improper prior is:

$$\hat{\sigma}^2 = e'(I + \tau R)e \,/\, (n - p). \tag{7.4}$$

The estimate (7.4) is quite similar to Wahba's estimate (7.1); (7.4) modifies (7.1) by including $(I + \tau R)$ in both the numerator and the denominator to obtain a weighted sum of squares divided by a constant divisor. In addition, (7.4) is precisely the estimate of σ^2 that would result from a conventional generalized least squares analysis of the regression model $Y = X\beta + \varepsilon$ when the errors are correlated with $E\{\varepsilon\varepsilon'\} = \sigma^2(I + \tau R)$.

Another possibility, again suggested by Lindley and Smith (1972), is to estimate σ^2 by the mode of the joint posterior of σ^2 and β, rather than the mode of the marginal posterior of σ^2. If we denote the prior densities of σ^2 and β by $p(\sigma^2)$ and $p(\beta)$, respectively, then their joint posterior will be:

$$p(\sigma^2,\beta/Y) \propto \frac{p(\sigma^2)p(\beta)\ \exp\{-(Y-X\beta)'(I+\tau R)^{-1}(Y-X\beta)/2\sigma^2\}}{\{\det(I+\tau R)\}^{1/2}\ \sigma^n}. \tag{7.5}$$

For each fixed value of σ^2, the posterior distribution of β is normal with mean (and modal) value $[X'(I+\tau R)^{-1}X]^{-1}X'(I+\tau R)^{-1}Y$. To find the joint modes, then, we need only substitute the modal value of β into (7.5) and maximize with respect to σ^2. It is easy to

verify that, with this substitution, the quadratic form in the numerator of (7.5) is equal to $\mathbf{e}'(\mathbf{I}+\tau\mathbf{R})\mathbf{e}$. If, a priori, σ^2 is assigned an inverse chi-square distribution, then the resulting function of σ^2 will retain the form of an inverse chi-square density. A special limiting case of the inverse chi-square is the commonly used improper prior in which $p(\sigma^2) \propto 1/\sigma^2$, for which (7.5) leads to the modal estimate:

$$\hat{\sigma}^2 = \mathbf{e}'(\mathbf{I}+\tau\mathbf{R})\mathbf{e} \;/\; (n + 2). \tag{7.6}$$

If we were to consider (7.5) evaluated at the mode of β as an inverse chi-square density for σ^2, the corresponding mean estimate would be:

$$\hat{\sigma}^2 = \mathbf{e}'(\mathbf{I}+\tau\mathbf{R})\mathbf{e} \;/\; (n - 2). \tag{7.7}$$

Wecker and Ansley (1983) analyzed the cubic spline model from a frequentist, rather than a Bayesian, perspective and derived a maximum likelihood estimate of σ^2:

$$\hat{\sigma}^2 = \mathbf{e}'(\mathbf{I}+\tau\mathbf{R})\mathbf{e} \;/\; n. \tag{7.8}$$

The derivation of (7.8) is identical to that of (7.6) except that no prior is introduced, so that only the likelihood is maximized.

7.2 Estimating τ

We have already seen that the estimated response function and the precision of the estimates can vary considerably as functions of τ. Thus it is clearly important to consider methods for estimating τ. The spline literature, in particular, has devoted considerable attention to this problem. Craven and Wahba (1977)

proposed choosing τ to minimize the generalized cross validation (GCV) function:

$$V(\tau) = \mathbf{e}'\mathbf{e} \;/\; \left[\mathrm{tr}(\mathbf{B})\right]^2, \tag{7.9}$$

where $\mathbf{e}$ and $\mathbf{B}$ are defined as in Section 7.1. Since $V(\tau)$ is not a simple function of τ, a numerical search algorithm is necessary to find the estimate. The GCV method is justified as an approximation to choosing τ to minimize mean squared error and has been found to work well provided the data are not too sparse.

Wecker and Ansley (1983) proposed the use of maximum likelihood to estimate τ, which amounts to choosing τ to minimize:

$$\mathbf{e}'(\mathbf{I}+\tau\mathbf{R})\mathbf{e} \;/\; \left\{\det\left[(\mathbf{I}+\tau\mathbf{R})^{-1}\right]\right\}^{1/n}. \tag{7.10}$$

As with the GCV method, a numerical search is necessary to carry out the minimization.

A strictly Bayesian approach would be to postulate a prior distribution for τ and then to compute its posterior distribution given the observed data. Moreover, since we have derived posterior distributions for the response function that are conditional on τ, it would then be appropriate to derive the marginal distribution of the response function by averaging over the posterior of τ. Unfortunately, the posterior distribution will inevitably be quite complicated, so that the averaging with respect to τ that is contemplated above will probably be intractable analytically. We could, as an approximation, simply derive a point estimate of τ and then proceed to estimate the response function as though that were the true value, a procedure that is reminiscent of empirical

Bayes estimation (see, for example, Morris 1983). The point estimate for τ, like that for σ^2, could be derived as the posterior mode and, assuming that the posterior is dominated by the likelihood, would be similar to Wecker and Ansley's maximum likelihood estimate (7.10).

It is not at all clear what would constitute a reasonable prior distribution for τ. Ideally, a prior for τ should reflect the experimenter's beliefs about how severe the bias is likely to be relative to experimental error. Young (1977) suggested that the prior for τ be chosen from the family of inverse chi-square distributions, but it is still necessary to specify a prior mean and variance for τ. Another possibility would be to assign an "uninformative" prior to τ, an approach that is often advocated for parameters about which prior information is weak. But what should this uninformative prior be? One candidate is $p(\tau) \propto 1/\tau$, since τ is a scale parameter in the model. (This would also be the uninformative prior from within the inverse chi-square family.) But the model here is a complicated one and, without further research, it is not clear what priors will lead to reasonable answers.

7.3 Estimating $R(\mathbf{u},\mathbf{v})$

The covariance function $R(\mathbf{u},\mathbf{v})$ may be seen here as a high-dimensional nuisance parameter in which we have little or no intrinsic interest but which we require in order to estimate the

response function. Although non-parametric estimation of $R(u,v)$ might be possible, the common strategy has been to propose some simple parametric form such as those described in Section 3. The problem of estimating R is thus reduced to one of estimating the associated parameters, which could be accomplished using the methods described in Section 7.2 for estimating τ. The parameters in R, however, typically enter the likelihood in a much more complicated fashion than does τ, so that all the comments in the preceding section on the difficulty of estimating τ and of assigning it a prior distribution will be even more true of parameters in R.

8. EXAMPLES

In this section we present two examples to illustrate the Bayesian response surface model. The first example involves simulated data with just one explanatory variable; the second example considers actual data from a chemical response surface experiment with two explanatory variables. All the Bayes estimates described in this section were calculated using the MATLAB matrix laboratory package (see Moler 1981).

8.1 Simulated Data

Suppose the true mean response to a scaled, centered explanatory variable x is given by the response function:

$$g(x) = x + 3x/(1 + x + 3x^2), \qquad (8.1)$$

which is the sum of a linear function and an inverse polynomial (see Nelder 1966 for a discussion of inverse polynomials). The graph of this function appears in Figure 1. Experimental data were simulated by adding computer generated random errors to 15 equally spaced design points between $x = -1.4$ and $x = 1.4$. The random errors were generated from a normal$(0,0.2^2)$ distribution via the NRANDOM command in Minitab (see Ryan, Joiner, and Ryan 1981). The simulated data appear as asterisks in Figure 1.

To model the data, we used a straight line as a graduating function (assigning improper priors to its coefficients) and defined the bias term by considering an expansion of the response function in normalized Hermite polynomials (see Steinberg 1984 for

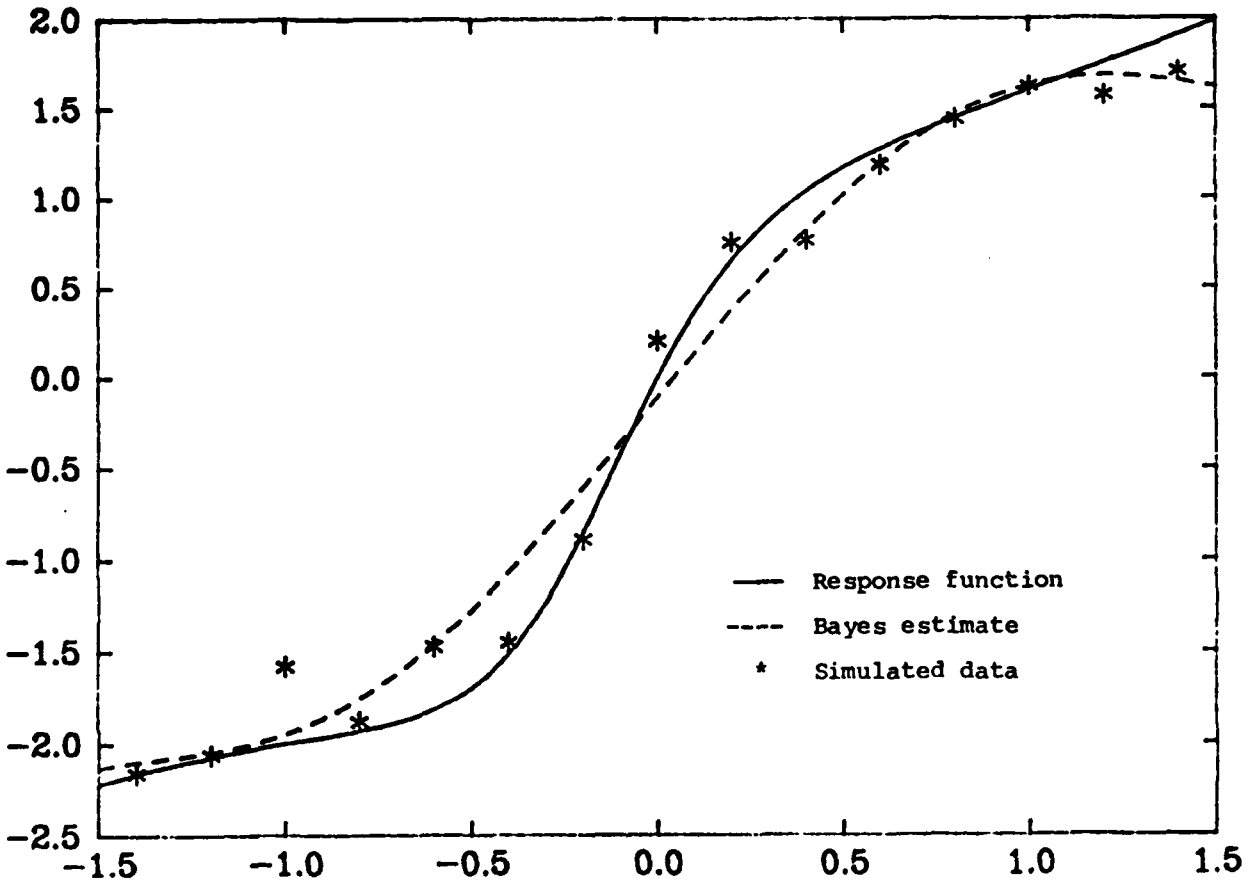


Figure 1: Bayes estimate of the response function $g(x) = x + 3x/(1 + x + 3x^2)$ from simulated data.

details). The resulting covariance function for the bias (2.3e) is:

$$\mathrm{Cov}\{\eta_u, \eta_v\} = \tau\sigma^2 \frac{\exp\{-(u-v)^2 w^2/(1-w^2) + 2uvw/(1-w)\}}{(1-w^2)^{1/2}} \qquad (8.2)$$

where w is a smoothing parameter that indicates the rate at which higher degree polynomials in the expansion are downweighted. Rather than estimate w, we chose to set $w = 0.4$; similar results were obtained with other values of w. We used generalized cross validation (7.9) to estimate τ; the criterion function leveled off at about $\tau = 5$ and we used this as our estimate of τ. (The maximum likelihood criterion (7.10) indicated that τ should be much closer to 0 and, in both this and the following example, appeared to severely underestimate τ.) Both (7.1) and (7.4) were considered as estimators of σ^2 and, for the estimated value of τ, they were in near agreement, giving estimates of .064 and .060, respectively (for other values of τ, (7.1) and (7.4) differed by as much as 30%).

The Bayes estimate of $g(x)$ is graphed in Figure 1. It does an excellent job of reproducing g for low and high values of x; although it is less successful in the middle region, this seems in large part to be the consequence of the error sequence, for which most of the errors at low x's were positive but most of the errors at high x's were negative. In particular, the large positive error term at $x = -0.4$ and the large negative error at $x = 0.4$ appear to strongly influence the fit in the middle of Figure 1. The

Bayes estimate certainly matches both the data and the true response function far more closely than would the OLS straight line fit here, and we think it is a useful alternative.

Figure 2 contrasts the variance of the estimated response function under the Bayesian model and under OLS. Since the variances are symmetric about 0, Figure 2 includes only positive x. Three features are noteworthy. First, for fixed σ^2, the Bayesian model with bias always suggests larger variances than does OLS. Over most of the experimental range, the variances are two to three times larger with the Bayesian model. Second, for small x, the Bayesian variances actually increase more slowly than the OLS variances. This phenomenon can be explained by remembering that, with the Bayesian model, most of the information used to estimate g(x) comes from nearby observations. Thus g(1/2) can be estimated with almost as much precision as g(0) because the distribution of design points close to 1/2 is about the same as that close to 0. Third, for x's near the edge of and outside the range of the data, variances increase dramatically with the Bayesian model. Reasonably precise estimates are possible only over the range of the data, where we have hard information on the nature of the response function; outside that range, precision is markedly worse. The Bayesian model, by explicitly stating that a tentative model will be subject to bias, leads to the realistic, albeit pessimistic, conclusion that the data provide little basis for extrapolation.

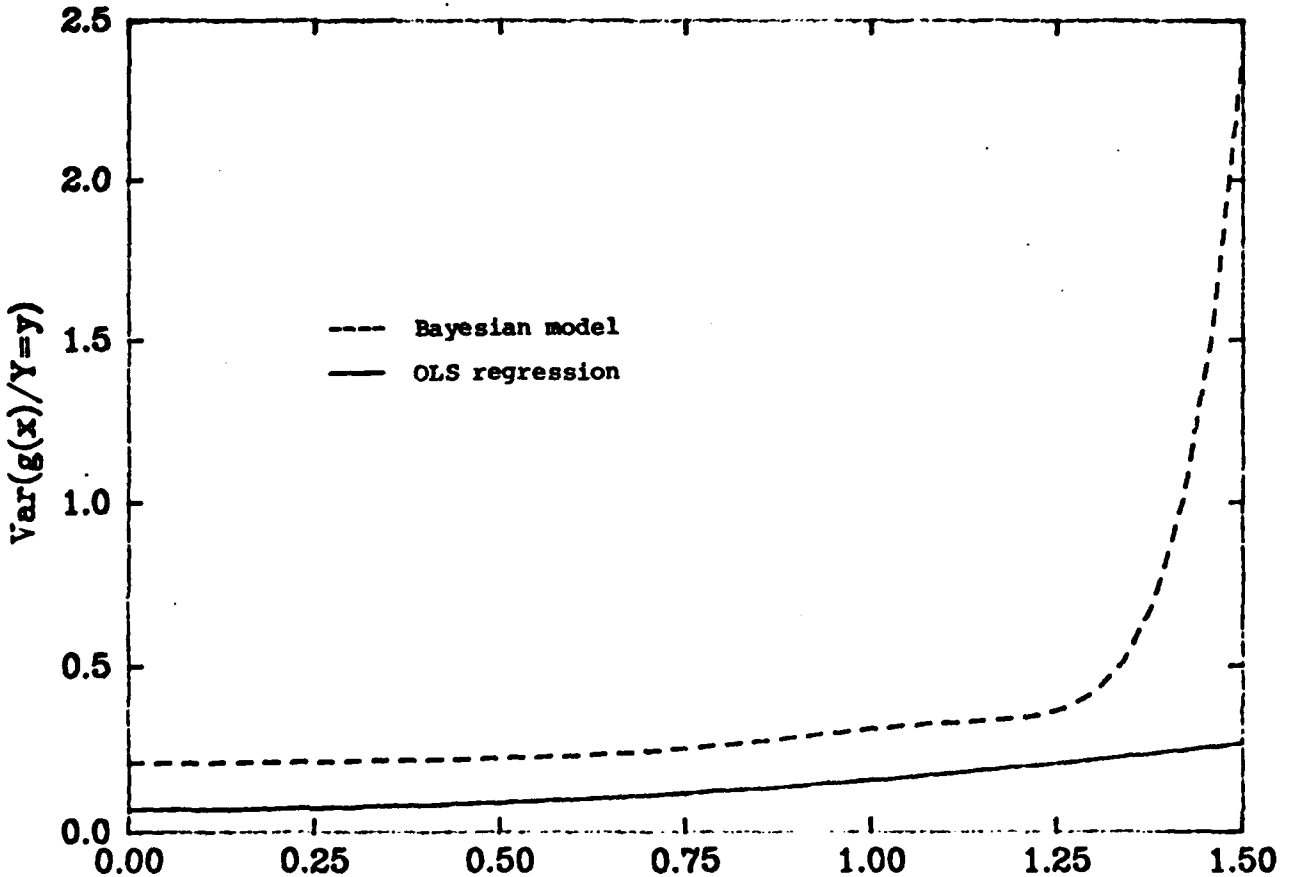


Figure 2: Variance of the estimated responses using the Bayesian model and using OLS regression. (The variances are scaled, for convenience, by $\sigma^2=1$.)

8.2 Chemical Experiment

Myers (1976) discusses a response surface experiment to model yield of the chemical mercaptobenzothiazole (MBT) as a function of reaction time and temperature. The experiment used a rotatable central composite design which is commonly used for fitting second order polynomials. The observed data are listed in Table 1 and shown in Figure 3. The number at the center of the design in Figure 3 is the average of the three runs made there; also, there was an additional run made at (-1,0) which we have deleted from our analysis in order to preserve the central composite structure (our results would not change much if the run were included).

Following Myers, we fit a second order polynomial in time and temperature to the MBT data using OLS. Figure 4 shows a contour plot of estimated yields from the OLS fit. The duplicate runs at the origin make possible a standard test for lack of fit of the second order polynomial model. The mean square error for lack of fit is 39.61 with 3 d.f. compared with a pure error mean square of 0.763 with 2 d.f. from the center replicates, indicating highly significant lack of fit. Inspection of the data (as well as any standard diagnostics) reveals that the two low yields in the northeast corner of Figure 3 are the major source of discrepancy. One possible explanation is that these two observations had unusual errors, but their proximity suggests, to the contrary, that they accurately reflect a severe degradation of yield when temperature and reaction time are too high. The OLS fit is unable to account

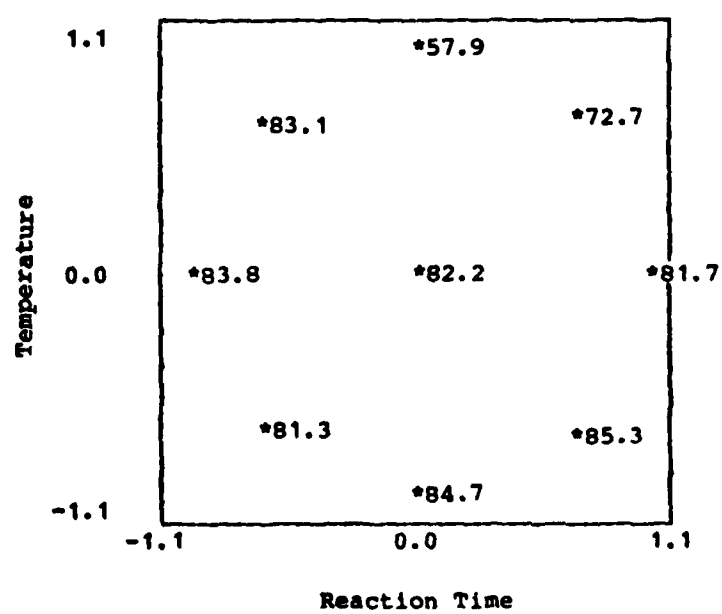
Table 1: Data from the MBT experiment. The explanatory variables have been standardized so that

$$X_1 = \frac{\text{Reaction time} - 12 \text{ minutes}}{8 \text{ minutes}}$$

$$X_2 = \frac{\text{Temperature} - 250 \text{ degrees C}}{30 \text{ degrees C}}$$

X_1	X_2	Y
-0.71	-0.71	81.3
0.71	-0.71	85.3
-0.71	0.71	83.1
0.71	0.71	72.7
-1.00	0.00	83.8
1.00	0.00	81.7
0.00	-1.00	84.7
0.00	1.00	57.9
0.00	0.00	82.4
0.00	0.00	82.9
0.00	0.00	81.2

Figure 3: Plot of the MBT data.



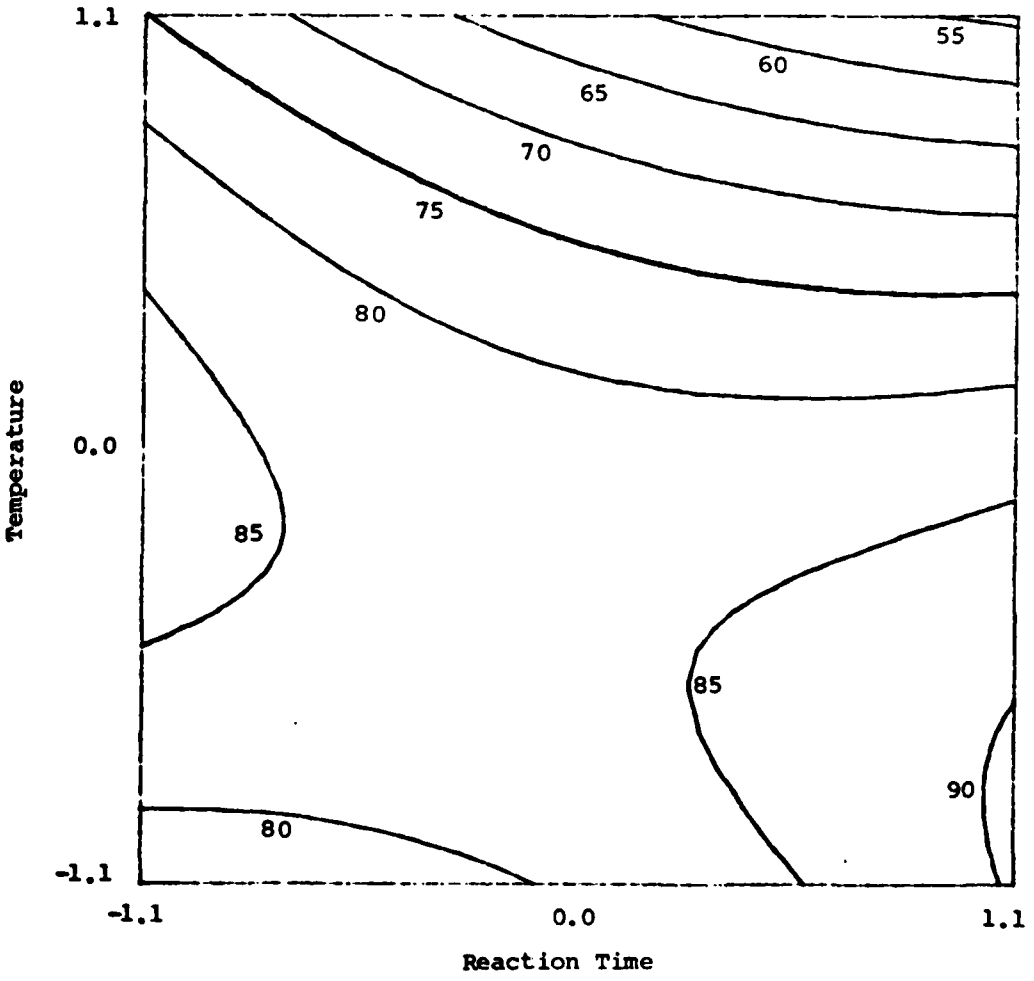


Figure 4: Contour plot of estimated MBT yield based on OLS regression.

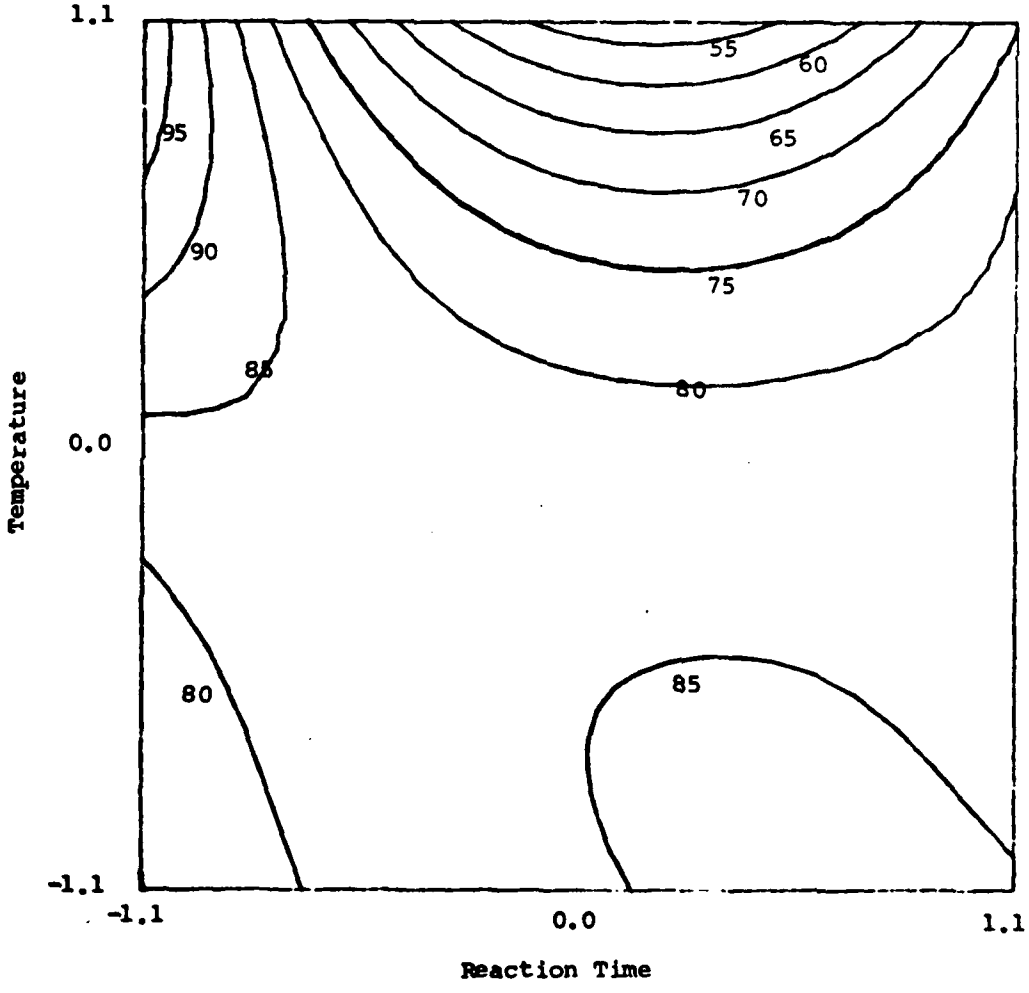


Figure 5: Contour plot of Bayes estimates of MBT yield.

for this sharp degradation and, as a result, all the OLS estimates are biased.

Now consider a Bayesian model for the MBT data that accounts for the possible presence of bias. We used a second order polynomial in time and temperature as the graduating function and defined the bias covariance to correspond to a two-dimensional expansion using Hermite polynomials (see Steinberg 1984 for details). The resulting covariance function is:

$$\mathrm{Cov}\{\eta_{\mathbf{u}},\eta_{\mathbf{v}}\} = \tau\sigma^2 \frac{\exp\{-(\mathbf{u}-\mathbf{v})'(\mathbf{u}-\mathbf{v})w^2/(1-w^2) + 2\mathbf{u}'\mathbf{v}w/(1-w)\}}{1-w^2} . \tag{8.3}$$

As in the first example, w is a smoothing parameter that indicates how rapidly higher-degree terms should be discounted. Similar results were obtained for a range of values of w and we elected to set $w = 0.4$ rather than attempting to estimate it from the data. For this choice of w, the generalized cross validation (7.9) estimate of τ was $\tau = 80$. To estimate σ^2, we computed both (7.1) and a modified version of (7.4), with the pure error sum of squares subtracted from the numerator and the associated degrees of freedom subtracted from the denominator. The resulting estimates were 0.765 and 0.756, respectively, in close agreement with the pure error estimate of 0.763.

A contour plot of the Bayes estimates of MBT yield is given in Figure 5. The Bayes estimates differ from the OLS estimates (in Figure 4) most noticeably across the top of the plots. The Bayes

estimates are high in the northwest corner of the plot and decrease rapidly as reaction time is increased, as indicated by the closeness of the contour lines in this region. The OLS estimates, on the other hand, are much higher in the immediate vicinity of the low observations and change gradually over the top half of the plot. The Bayes estimates in the northeast corner of the plot suggest that yield will increase if reaction time is increased still further, a conclusion that seems implausible. The poor performance of the Bayes estimates in this region is really not surprising: it simply reflects the inability of any model to give accurate estimates throughout the range of time and temperature when there are so few degrees of freedom to estimate the bias component. Also, it is worth remembering that this region is situated outside the circular design region, so the Bayes estimates there have very large variances. Within the design region, we think the Bayes estimates are superior to the OLS estimates.

9. DISCUSSION

In this paper, we have analyzed a Bayesian model for estimating response surfaces. The key feature of the Bayesian model is the inclusion of a term that explicitly represents the bias that arises when a simple graduating function is used to provide a local approximation to a complex response function. The resulting estimate of the response function includes an estimate of the

graduating function and a "correction for bias" term that lends it the flexibility to accurately represent data when the graduating function alone is inadequate. We have derived measures of precision for the Bayes response surface estimates and have shown that OLS may lead to an overly optimistic assessment of precision if bias is present.

The bias term in the model is described in terms of a prior probability distribution and we have emphasized how the nature of that distribution, in particular its covariance function (2.3e), determine the nature of the estimate of the response surface. We have reviewed suggestions in the literature for defining realistic covariance functions and methods that have been proposed for estimating covariance parameters. We believe that there is much room for useful research on this topic.

In building empirical models, we usually seek some compromise between the goals of good fit and model simplicity. Occasionally, we find ourselves in the ideal situation where good fit can be achieved with a simple model, but such models often prove too rigid to provide a good fit. We think that the Bayesian model discussed here offers a useful approach for introducing additional flexibility in empirical modeling.

APPENDIX

This section contains proofs for some of the theorems stated in the text.

Theorem 4.1: Suppose $\mathbf{X}$ has full column rank and an improper prior is assigned to the regression coefficients in (2.2)-(2.3). Then:

$$\hat{g}(\mathbf{x}) = \lim_{\mathbf{V}^{-1} \to \mathbf{0}} E\{g(\mathbf{x})/\mathbf{Y}=\mathbf{y}\} \tag{4.1}$$
$$= \mathbf{f}'(\mathbf{x})\left[\mathbf{X}'\mathbf{M}^{-1}\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{M}^{-1}\mathbf{y} +$$
$$\tau\sigma^2\mathbf{r}'(\mathbf{x})\left\{\mathbf{M}^{-1} - \mathbf{M}^{-1}\mathbf{X}\left[\mathbf{X}'\mathbf{M}^{-1}\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{M}^{-1}\right\}\mathbf{y},$$

where $\mathbf{M} = (\mathbf{I} + \tau\mathbf{R})$, and

$$\lim_{\mathbf{V}^{-1} \to \mathbf{0}} E\{\beta/\mathbf{Y}=\mathbf{y}\} = \left[\mathbf{X}'\mathbf{M}^{-1}\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{M}^{-1}\mathbf{y}.$$

Proof: From Theorem 3.1,

$$E\{g(\mathbf{x})/\mathbf{Y}=\mathbf{y}\} = \mathbf{f}'(\mathbf{x})\beta_0 + \left[\tau\sigma^2\mathbf{r}'(\mathbf{x}) + \mathbf{f}'(\mathbf{x})\mathbf{V}\mathbf{X}'\right]$$
$$\times \left(\sigma^2\mathbf{I} + \tau\sigma^2\mathbf{R} + \mathbf{X}\mathbf{V}\mathbf{X}'\right)^{-1}(\mathbf{y} - \mathbf{X}\beta_0)$$
$$= \mathbf{f}'(\mathbf{x})\beta_0 + \mathbf{f}'(\mathbf{x})\mathbf{V}\mathbf{X}'\left(\sigma^2\mathbf{I} + \tau\sigma^2\mathbf{R} + \mathbf{X}\mathbf{V}\mathbf{X}'\right)^{-1}(\mathbf{y} - \mathbf{X}\beta_0)$$
$$+ \tau\sigma^2\mathbf{r}'(\mathbf{x})\left(\sigma^2\mathbf{I} + \tau\sigma^2\mathbf{R} + \mathbf{X}\mathbf{V}\mathbf{X}'\right)^{-1}(\mathbf{y} - \mathbf{X}\beta_0)$$

and applying Lemma 2 to the first line and Lemma 1 to the second line of the last expression,

$$= \mathbf{f}'(\mathbf{x})\beta_0 + \mathbf{f}'(\mathbf{x})\left[\sigma^{-2}\mathbf{X}'\mathbf{M}^{-1}\mathbf{X} + \mathbf{V}^{-1}\right]^{-1}\sigma^{-2}\mathbf{X}'\mathbf{M}^{-1}(\mathbf{y}-\mathbf{X}\beta_0)$$
$$+ \tau\sigma^2\mathbf{r}'(\mathbf{x})\left\{\sigma^{-2}\mathbf{M}^{-1} - \sigma^{-2}\mathbf{M}^{-1}\mathbf{X}\left[\sigma^{-2}\mathbf{X}'\mathbf{M}^{-1}\mathbf{X} + \mathbf{V}^{-1}\right]^{-1}\right.$$
$$\left.\times\, \sigma^{-2}\mathbf{X}'\mathbf{M}^{-1}\right\}(\mathbf{y}-\mathbf{X}\beta_0)$$

and provided $\mathbf{X}$ has full column rank, as $\mathbf{V}^{-1} \to \mathbf{0}$, this converges to:

$$\mathbf{f}'(\mathbf{x})\beta_0 + \mathbf{f}'(\mathbf{x})\left[\mathbf{X}'\mathbf{M}^{-1}\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{M}^{-1}(\mathbf{y}-\mathbf{X}\beta_0)$$
$$+ \tau\mathbf{r}'(\mathbf{x})\left\{\mathbf{M}^{-1} - \mathbf{M}^{-1}\mathbf{X}\left[\mathbf{X}'\mathbf{M}^{-1}\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{M}^{-1}\right\}(\mathbf{y}-\mathbf{X}\beta_0)$$

$$= f'(x)\left[X'M^{-1}X\right]^{-1}X'M^{-1}y$$
$$+ \tau r'(x)\left\{M^{-1} - M^{-1}X\left[X'M^{-1}X\right]^{-1}X'M^{-1}\right\}y.$$

The proof of (4.2) is contained in the proof of (4.1).

Theorem 5.1: Given the model (2.2)-(2.3), suppose that the matrix $\bar{R}$ derived from R by eliminating identical rows and columns is non-singular. Then:

$$\hat{Y}_i = \text{average of all observations at } x_i. \qquad (5.2)$$

Proof: Denote the distinct design points by $x_1, \ldots, x_m$, and denote by $\bar{Y}$ the $mx1$ vector of average responses at the design points. The sampling density of Y can be factored into the density of $\bar{Y}$ times the density of Y given $\bar{Y}$. The latter is clearly independent of $g(x)$, so we can compute the posterior distribution of $g(x)$ by conditioning only on $\bar{Y}$. The sampling distribution of $\bar{Y}$, conditional on β and η, is:

$$\bar{Y} \sim N(X\beta + \eta, \sigma^2 D),$$

where X is the mxp matrix whose ith row is $f'(x_i)$, η is the $mx1$ vector of bias terms at the distinct design points, and D is a diagonal matrix whose ith entry is $1/n_i$, where n_i is the number of observations at x_i. The prior covariance matrix of η is $\tau\sigma^2\bar{R}$, where $\bar{R}$ is precisely the matrix obtained by eliminating duplicate rows and columns of R. If we recompute (3.2), (4.1), and (4.9) conditioning on $\bar{Y}$, we obtain (5.2) directly from (4.9) under the assumption that $\bar{R}$ is non-singular.

Theorem 5.3: $\hat{\mathbf{Y}}$ solves the minimization problem: find $\mathbf{u}$ to minimize

$$(\mathbf{u}-\mathbf{y})'(\mathbf{u}-\mathbf{y}) + (\mathbf{u}-\mathbf{X}\boldsymbol{\beta}_0)'\sigma^2(\tau\sigma^2\mathbf{R} + \mathbf{X}\mathbf{V}\mathbf{X}')^{-1}(\mathbf{u}-\mathbf{X}\boldsymbol{\beta}_0). \qquad (5.5)$$

If $\mathbf{R}$ is non-singular, then (5.5) converges to:

$$(\mathbf{u}-\mathbf{y})'(\mathbf{u}-\mathbf{y}) + \tau^{-1}\mathbf{u}'\left[\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}^{-1}\right]\mathbf{u} \qquad (5.6)$$

as $\mathbf{V}^{-1} \to \mathbf{0}$. The second term in (5.6) is 0 if and only if $\mathbf{u} \in \text{col}(\mathbf{X})$.

Proof: Denote by $\boldsymbol{\theta}$ the vector of expected values at the design points, as in (2.4). Then, by definition, $\hat{\mathbf{Y}}$ is the posterior expectation of $\boldsymbol{\theta}$. From (2.4), the prior distribution of $\boldsymbol{\theta}$ is:

$$\boldsymbol{\theta} \sim N(\mathbf{X}\boldsymbol{\beta}_0, \tau\sigma^2\mathbf{R} + \mathbf{X}\mathbf{V}\mathbf{X}'),$$

and the sampling distribution of $\mathbf{Y}$ given $\boldsymbol{\theta}$ is:

$$\mathbf{Y}/\boldsymbol{\theta} \sim N(\boldsymbol{\theta}, \sigma^2\mathbf{I}).$$

Applying Bayes' Theorem, the posterior density of $\boldsymbol{\theta}$ is proportional to:

$$\exp\left\{-\left[\sigma^{-2}(\mathbf{y}-\boldsymbol{\theta})'(\mathbf{y}-\boldsymbol{\theta}) + (\boldsymbol{\theta}-\mathbf{X}\boldsymbol{\beta}_0)'(\tau\sigma^2\mathbf{R} + \mathbf{X}\mathbf{V}\mathbf{X}')^{-1}(\boldsymbol{\theta}-\mathbf{X}\boldsymbol{\beta}_0)\right]/2\right\}$$

$$= \exp\left\{-Q(\boldsymbol{\theta})/2\right\},$$

which corresponds to the density for a normal distribution. Thus the posterior mean of $\boldsymbol{\theta}$, $\hat{\mathbf{Y}}$, can be found by minimizing the quadratic form $Q(\boldsymbol{\theta})$, and that is clearly equivalent to minimizing (5.5). We can easily derive (5.6) from (5.5) by applying Lemma 1 of Section 4 with $\mathbf{A} = \tau\sigma^2\mathbf{R}$. If $\mathbf{u} \in \text{col}(\mathbf{X})$, then $\mathbf{u} = \mathbf{X}\boldsymbol{\gamma}$ for some vector $\boldsymbol{\gamma}$ and it is easy to verify that the second term in (5.6) is 0. To prove the converse, note that the second term has the form $\mathbf{u}'\mathbf{L}\mathbf{u}$, where the matrix $\mathbf{L}$ is obtained as the limit of the matrix

in the second term of (5.5). This latter matrix is positive definite, so $\mathbf{L}$ must be positive semi-definite. Thus $\mathbf{u'Lu} = 0$ implies $\mathbf{Lu} = \mathbf{0}$ and $\mathbf{0} = \mathbf{RLu} = \mathbf{Pu}$, where $\mathbf{P} = \mathbf{RL} = \mathbf{I} - \mathbf{X}(\mathbf{X'R}^{-1}\mathbf{X})^{-1}\mathbf{X'R}^{-1}$ is an idempotent matrix that projects into the orthogonal complement of $\text{col}(\mathbf{X})$. Thus $\mathbf{Pu} = \mathbf{0}$ implies that $\mathbf{u} \in \text{col}(\mathbf{X})$, as claimed.

Theorem 6.2: Given (2.2)-(2.3),

$$\lim_{\mathbf{V}^{-1} \to \mathbf{0}} \text{Var}\{g(x)/\mathbf{Y}=\mathbf{y}\} = \sigma^2\{\tau R(x,x) + \mathbf{f}'(x)(\mathbf{X'M}^{-1}\mathbf{X})^{-1}\mathbf{f}(x) - 2\tau\mathbf{r}'(x)\mathbf{M}^{-1}\mathbf{X}(\mathbf{X'M}^{-1}\mathbf{X})^{-1}\mathbf{f}(x) - \tau^2\mathbf{r}'(x)[\mathbf{M}^{-1} - \mathbf{M}^{-1}\mathbf{X}(\mathbf{X'M}^{-1}\mathbf{X})^{-1}\mathbf{X'M}^{-1}]\mathbf{r}(x)\}. \tag{6.4}$$

$$\lim_{\mathbf{V}^{-1} \to \mathbf{0}} \text{Var}\{\boldsymbol{\beta}/\mathbf{Y}=\mathbf{y}\} = \sigma^2(\mathbf{X'M}^{-1}\mathbf{X})^{-1}. \tag{6.5}$$

Proof: Applying Lemma 1 of Section 4 in reverse to (6.2) yields the identity:

$$\mathbf{V} - \mathbf{VX}'(\sigma^2\mathbf{I} + \tau\sigma^2\mathbf{R} + \mathbf{XVX}')^{-1}\mathbf{XV} = [\sigma^{-2}\mathbf{X'M}^{-1}\mathbf{X} + \mathbf{V}^{-1}]^{-1}$$

and (6.5) follows immediately. To prove (6.4), note that we can rewrite (6.1) as:

$$\text{Var}\{g(x)/\mathbf{Y}=\mathbf{y}\} = \tau\sigma^2 R(x,x) + \mathbf{f}'(x)[\mathbf{V} - \mathbf{VX}'(\sigma^2\mathbf{M} + \mathbf{XVX}')^{-1}\mathbf{XV}]\mathbf{f}(x) - 2\tau\sigma^2\mathbf{r}'(x)[\sigma^2\mathbf{M} + \mathbf{XVX}']^{-1}\mathbf{XVf}(x) - \tau^2\sigma^4\mathbf{r}'(x)[\sigma^2\mathbf{M} + \mathbf{XVX}']^{-1}\mathbf{r}(x).$$

Now apply the above identity to the first line, Lemma 2 of Section 4 to the second line, and Lemma 1 of Section 4 to the third line. Taking limits as $\mathbf{V}^{-1} \to \mathbf{0}$ yields (6.4).

Theorem 6.3: Given the model (2.2)-(2.3):

(i) The posterior variance of $g(\mathbf{x})$ is a monotone increasing function of τ.

(ii) The posterior variance of $g(\mathbf{x})$ obtains a minimum value of

$$\sigma^2 \mathbf{f}'(\mathbf{x})(\mathbf{X}'\mathbf{X} + \sigma^2\mathbf{V}^{-1})^{-1}\mathbf{f}(\mathbf{x})$$

when $\tau = 0$. If the regression coefficients are assigned an improper prior, the minimum value is

$$\sigma^2 \mathbf{f}'(\mathbf{x})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{f}(\mathbf{x}).$$

(iii) If $\mathbf{x}$ is a design point, then:

$$\mathrm{Var}\{\hat{g}(\mathbf{x})/\mathbf{Y}=\mathbf{y}\} \leq \sigma^2.$$

(iv) If both $\bar{\mathbf{R}}$ and $\bar{\mathbf{R}}_{\mathbf{x}}$ are non-singular, then the posterior variance of $g(\mathbf{x})$ diverges to infinity as $\tau \to \infty$.

(v) The posterior variance of $\boldsymbol{\beta}$ is a monotone increasing function of τ.

(vi) The minimum posterior variance of $\boldsymbol{\beta}$ is attained when $\tau = 0$ and is

$$\sigma^2(\mathbf{X}'\mathbf{X} + \sigma^2\mathbf{V}^{-1})^{-1}.$$

If the regression coefficients are assigned an improper prior, the minimum value is

$$\sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

Proof: Results (ii) and (vi) are trivial. To prove (i), denote by $\mathbf{W}(\tau)$ the prior covariance matrix of $(\mathbf{Y}', g(\mathbf{x}))$. This matrix has the form: $\mathbf{W}(\tau) = \sigma^2\mathbf{I}^* + \tau\sigma^2\mathbf{H}^* + \mathbf{X}^*\mathbf{V}\mathbf{X}^{*\prime}$. The posterior variance of $g(\mathbf{x})$ is simply the inverse of the lower right-hand corner element of $\mathbf{W}(\tau)^{-1}$; that is,

$$\mathrm{Var}\{g(\mathbf{x})/\mathbf{Y}=\mathbf{y}\} = \left[\mathbf{v}_{n+1}'\mathbf{W}(\tau)^{-1}\mathbf{v}_{n+1}\right]^{-1},$$

where $\mathbf{v}_{n+1}$ is a unit vector whose $(n+1)$st element is 1 and whose remaining elements are 0. Now, suppose $\tau_2 > \tau_1 > 0$. Then $\mathbf{W}(\tau_2) - \mathbf{W}(\tau_1)$ will be positive semi-definite, because $\mathbf{R}^*$ is positive semi-definite, and $\mathbf{W}(\tau_1)^{-1} - \mathbf{W}(\tau_2)^{-1}$ will also be positive semi-definite. Therefore:

$$\begin{aligned} 0 &\leq \mathbf{v}_{n+1}'\left[\mathbf{W}(\tau_1)^{-1} - \mathbf{W}(\tau_2)^{-1}\right]\mathbf{v}_{n+1} \\ &= \left[\mathrm{Var}\{g(\mathbf{x})/\mathbf{Y}=\mathbf{y},\tau_1\}\right]^{-1} - \left[\mathrm{Var}\{g(\mathbf{x})/\mathbf{Y}=\mathbf{y},\tau_2\}\right]^{-1} \end{aligned}$$

which implies that

$$\mathrm{Var}\{g(\mathbf{x})/\mathbf{Y}=\mathbf{y},\tau_1\} \leq \mathrm{Var}\{g(\mathbf{x})/\mathbf{Y}=\mathbf{y},\tau_2\},$$

proving the claimed monotonicity property.

(iii) It will suffice to consider the posterior variance at $\mathbf{x}_1$. Conditional on $g(\mathbf{x}_1)$, Y_1 has a normal $(g(\mathbf{x}_1),\sigma^2)$ distribution, so the conditional distribution of $g(\mathbf{x}_1)$ given Y_1, which we know to be normal, has a variance of at most σ^2. Thus:

$$\mathrm{Var}\{g(\mathbf{x}_1)/\mathbf{Y}=\mathbf{y}\} \leq \mathrm{Var}\{g(\mathbf{x}_1)/Y_1=y_1\} \leq \sigma^2.$$

(iv) As in Theorem 5.1, we can compute the posterior variance of $g(\mathbf{x})$ by conditioning only on the replicate averages, which yields a formula analogous to (6.1) but with $\bar{\mathbf{R}}$ in place of $\mathbf{R}$, an $m\times 1$ vector $\bar{\mathbf{r}}(\mathbf{x})$ in place of $\mathbf{r}(\mathbf{x})$, and a diagonal matrix $\mathbf{D}$ in place of $\mathbf{I}$, where $\mathbf{D}_{i,i} = 1/n_i$, and n_i is the number of observations at the ith design point. Rearranging terms, we obtain:

$$\begin{aligned} \mathrm{Var}\{g(\mathbf{x})/\mathbf{Y}=\mathbf{y}\} &= \mathbf{f}'(\mathbf{x})\left[\mathrm{Var}\{\beta/\mathbf{Y}=\mathbf{y}\}\right]\mathbf{f}(\mathbf{x}) \\ &\quad - 2\tau\sigma^2\bar{\mathbf{r}}'(\mathbf{x})\left[\sigma^2\mathbf{D} + \tau\sigma^2\bar{\mathbf{R}} + \mathbf{XVX}'\right]^{-1}\mathbf{XVf}(\mathbf{x}) \\ &\quad + \tau\sigma^2\mathbf{R}(\mathbf{x},\mathbf{x}) - \tau^2\sigma^4\bar{\mathbf{r}}'(\mathbf{x})\left[\sigma^2\mathbf{D} + \tau\sigma^2\bar{\mathbf{R}} + \mathbf{XVX}'\right]^{-1}\bar{\mathbf{r}}(\mathbf{x}). \end{aligned}$$

The first line of the last expression is clearly non-negative, regardless of the value of τ. The second term can be rewritten

$$- 2\sigma^2\bar{r}'(x)\left[\tau^{-1}\sigma^2\mathbf{D} + \sigma^2\bar{\mathbf{R}} + \tau^{-1}\mathbf{XVX}'\right]^{-1}\mathbf{XV}\bar{r}(x)$$

and converges to the finite limit $-2\bar{r}(x)\bar{\mathbf{R}}^{-1}\mathbf{XV}\bar{r}(x)$ as $\tau \to \infty$, when $\bar{\mathbf{R}}$ is non-singular. The final line can be rewritten

$$\tau\sigma^2\left\{R(x,x) - \sigma^2\bar{r}'(x)\left[\tau^{-1}\sigma^2\mathbf{D} + \sigma^2\bar{\mathbf{R}} + \tau^{-1}\mathbf{XVX}'\right]^{-1}\bar{r}(x)\right\}.$$

If $\bar{\mathbf{R}}$ is non-singular, the term in curly brackets converges to $R(x,x) - \bar{r}'(x)\bar{\mathbf{R}}^{-1}\bar{r}(x)$ as $\tau \to \infty$, and if $\bar{\mathbf{R}}_x$ is non-singular, this expression is positive. Thus, if both matrices are non-singular, the last line tends to infinity as $\tau \to \infty$, so that $\mathrm{Var}\{g(x)/\mathbf{Y}=\mathbf{y}\}$ diverges to infinity as $\tau \to \infty$.

(v) The proof of (v) follows the same lines as the proof of (i).

Theorem 6.4: The posterior variance matrix of $\boldsymbol{\theta}$ is:

$$\mathrm{Var}\{\boldsymbol{\theta}/\mathbf{Y}=\mathbf{y}\} = \sigma^2\mathbf{I} - \sigma^4(\sigma^2\mathbf{I}\ \tau\sigma^2\mathbf{R} + \mathbf{XVX}')^{-1}. \qquad (6.6)$$

(6.6) is monotone increasing in τ and achieves a minimum of $\sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X} + \sigma^2\mathbf{V}^{-1})^{-1}\mathbf{X}'$ when $\tau = 0$. If $\mathbf{R}$ is non-singular, (6.6) converges to $\sigma^2\mathbf{I}$ as $\tau \to \infty$.

$$\lim_{\mathbf{V}^{-1} \to \mathbf{0}} \mathrm{Var}\{\boldsymbol{\theta}/\mathbf{Y}=\mathbf{y}\} = \sigma^2\left\{\mathbf{I} - \mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{X}\left[\mathbf{X}'\mathbf{M}^{-1}\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{M}^{-1}\right\}. \qquad (6.7)$$

(6.7) is also monotone increasing in τ and achieves a minimum of $\sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ when $\tau = 0$. If $\mathbf{R}$ is non-singular, (6.7) converges to $\sigma^2\mathbf{I}$ as $\tau \to \infty$.

Proof: From (2.4), the joint distribution of $(\boldsymbol{\theta}', \mathbf{Y}')$ is multivariate normal with covariance matrix:

$$\frac{\tau\sigma^2\mathbf{R} + \mathbf{XVX}'}{\tau\sigma^2\mathbf{R} + \mathbf{XVX}'} \Bigg| \frac{\tau\sigma^2\mathbf{R} + \mathbf{XVX}'}{\sigma^2\mathbf{I} + \tau\sigma^2\mathbf{R} + \mathbf{XVX}'} .$$

From standard properties of multivariate normal distributions,

$$\mathrm{Var}\{\boldsymbol{\theta}/\mathbf{Y}=\mathbf{y}\} = \tau\sigma^2\mathbf{R} + \mathbf{XVX}' - (\tau\sigma^2\mathbf{R} + \mathbf{XVX}') \times (\sigma^2\mathbf{I} + \tau\sigma^2\mathbf{R} + \mathbf{XVX}')^{-1}(\tau\sigma^2\mathbf{R} + \mathbf{XVX}').$$

Now add and subtract $\sigma^2\mathbf{I}$ to each of the $(\tau\sigma^2\mathbf{R} + \mathbf{XVX}')$ terms. After some obvious cancellations, we obtain (6.6). Let $\mathbf{V}(\boldsymbol{\theta},\tau)$ denote (6.6) as a function of τ. If $\tau_2 > \tau_1$,

$$\mathbf{V}(\boldsymbol{\theta},\tau_2) - \mathbf{V}(\boldsymbol{\theta},\tau_1) = \sigma^4\left[\left(\sigma^2\mathbf{I} + \tau_1\sigma^2\mathbf{R} + \mathbf{XVX}'\right)^{-1} - \left(\sigma^2\mathbf{I} + \tau_2\sigma^2\mathbf{R} + \mathbf{XVX}'\right)^{-1}\right].$$

Were the matrices in the square brackets not inverted, their difference would clearly be negative semi-definite; with the inverses, then, the difference is positive semi-definite, as claimed. The minimum is then achieved when $\tau = 0$ and is

$$\mathbf{V}(\boldsymbol{\theta},0) = \sigma^2\mathbf{I} - \sigma^4(\sigma^2\mathbf{I} + \mathbf{XVX}')^{-1} = \sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X} + \sigma^2\mathbf{V}^{-1})^{-1}\mathbf{X}'$$

upon application of Lemma 1 of Section 4. The limit as $\tau \to \infty$ follows immediately from (6.6). We obtain (6.7) from (6.6) by applying Lemma 1 of Section 4 to the second term and taking limits as $\mathbf{V}^{-1} \to \mathbf{0}$. The monotonicity with respect to τ is unaltered by the limiting process and the minimum value for $\tau = 0$ follows from (6.7). To obtain the limit of (6.7) as $\tau \to \infty$, we first take limits of the expression derived above as $\mathbf{V}^{-1} \to \mathbf{0}$, obtaining $\sigma^2\mathbf{I} - \sigma^2\mathbf{B}(\tau)$, where $\mathbf{B}(\tau)$ is the matrix that maps $\mathbf{Y}$ into the residual vector. If $\mathbf{R}$ is non-singular, we know that the estimated

response function converges to an interpolant as $\tau \to \infty$, so $B(\tau)$ must converge to a 0 matrix. Thus, if R is non-singular, (6.7) converges to $\sigma^2 I$ as $\tau \to \infty$.

References

Anderson, T. (1958). An Introduction To Multivariate Statistical Analysis. New York: John Wiley and Sons, Inc.

Blight, B. J. N. and Ott, L. (1975). A Bayesian approach to model inadequacy for polynomial regression. Biometrika, 62, 79-88.

Box, G. E. P. (1954). The exploration and exploitation of response surfaces: Some general considerations and examples. Biometrics, 10, 16-60.

Box, G. E. P. and Draper, N. R. (1959). A basis for the selection of a response surface design. Jour. Amer. Stat. Assoc., 622-653.

Box, G. E. P. and Wilson, K. B. (1951). On the experimental attainment of optimum conditions. Jour. Roy. Stat. Soc., B, 13, 1-45.

Box, G. E. P. and Youle, P. V. (1955). The exploration and exploitation of response surfaces: An example of the link between the fitted surface and the basic mechanism of the system. Biometrics, 11, 287-323.

Craven, P. and Wahba, G. (1977). Smoothing noisy data with spline functions. Numerische Mathematik, 31, 377-403.

Kimeldorf, G. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. Jour. of Math. Anal. and Applic., 33, 82-95.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model. Jour. Roy. Stat. Soc., B, 34, 1-41.

Moler, C. (1981). MATLAB Users' Guide.

Morris, C. (1983). Parametric empirical Bayes inference: Theory and applications. Jour. Amer. Stat. Assoc., 78, 47-65.

Myers, R. H. (1976). Response Surface Methodology.

Nelder, J. A. (1966). Inverse polynomials, a useful group of multifactor response functions. Biometrics, 22, 128-141.

O'Hagan, A. (1978). Curve fitting and optimal design for prediction. Jour. Roy. Stat. Soc., B, 40, 1-41.

Ryan, T. A., Joiner, B. L. and Ryan, B. F. (1981). MINITAB Reference Manual.

Scheffé, H. (1947). A useful convergence theorem for probability distributions. Ann. Math. Stat., 18, 434-458.

Shepp, L. A. (1966). Radon-Nikodym derivatives of Gaussian measures. Ann. Math. Stat., 37, 321-354.

Smith, A. F. M. (1973). Bayes estimates in one-way and two-way models. Biometrika, 60, 319-329.

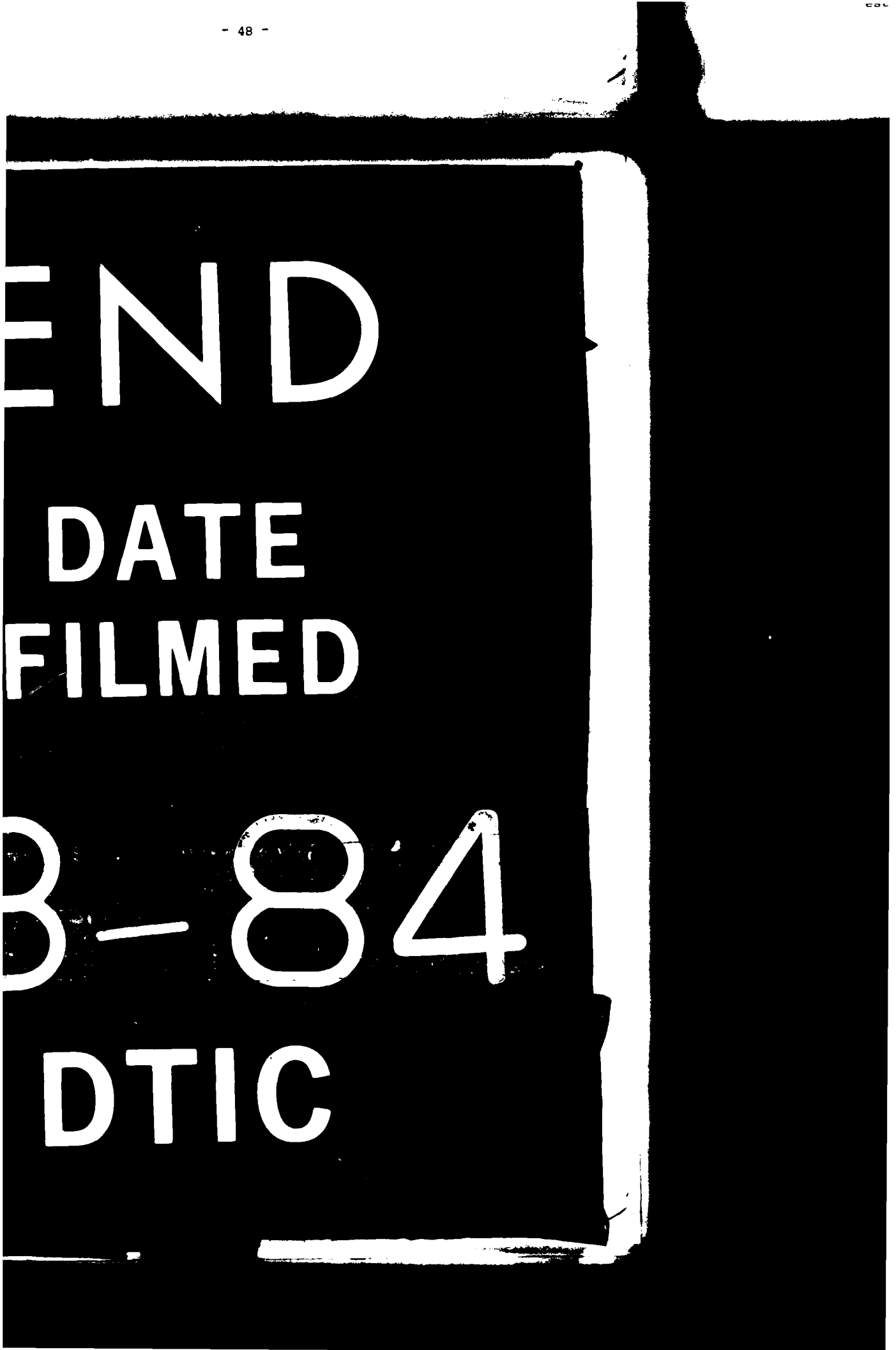
Steinberg, D. M. (1984). Bayesian models for response surfaces I: The equivalence of several models and their relationship to smoothing splines. Submitted for publication.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. Jour. Roy. Stat. Soc., B, 40, 364-372.

Wahba, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. Jour. Roy. Stat. Soc., B, 45, 133-150.

Wecker, W. E. and Ansley, C. F. (1983). The signal extraction approach to nonlinear regression and spline smoothing. Jour. Amer. Stat. Assoc., 78, 81-89.

Young, A. S. (1977). A Bayesian approach to prediction using polynomials. Biometrika, 64, 309-317.


END
DATE
FILMED
8-84
DTIC